Chapter 11
"Looking at" Educational Interventions: Surplus Value of a Complex Dynamic Systems Approach to Study the Effectiveness of a Science and Technology Educational Intervention

Sabine van Vondel, Henderien Steenbeek, Marijn van Dijk, and Paul van Geert

Introduction

There is no doubt that a classroom can be conceived of as a complex dynamic system, in that it consists of many interacting components —the students and the teacher— that influence each other's behavior and characteristics over a wide variety of nested time scales (Lewis, 2002; Smith & Thelen, 2003; Van Geert & Steenbeek, 2005). If one takes, for instance, a science lesson in a classroom consisting of 11-year-old students, then the teacher's questions during a science activity influence the reactions of the students. The interactions during this activity influence the interaction during the next activity or next lesson.

As this is an example of an educational system, the interactions at the behavioral level of the system are explicitly aimed at durably changing particular properties —such as the students' knowledge, skills, or understanding about science. Note that at the same time other properties, such as the order in the class or the level of involvement of the students, should be maintained. Modern schools that promote the lifelong learning of the teacher make decisions about programs for teacher professionalization, which are either reluctantly or enthusiastically received by the teachers (Wetzels, Steenbeek, & Van Geert, 2015). Such professional interventions are often presented as fixed protocols, but in reality they unfold as highly idiosyncratic processes. In fact they are emergent processes in which many components —including the written intervention protocol, the coach's capacities, the unique circumstances of the school and the time and effort invested by the teachers— are dynamically intertwined. Such interventions are in fact forms of perturbation in an existing, self-sustaining pattern of activities which takes place during real-time learning situations. Asking a particular type of questions, performing a particular

S. van Vondel (✉) • H. Steenbeek • M. van Dijk • P. van Geert
University of Groningen, Netherlands
e-mail: s.van.vondel@rug.nl

M. Koopmans, D. Stamovlasis (eds.), *Complex Dynamical Systems in Education*, DOI 10.1007/978-3-319-27577-2_11

type of activities or typical reactions of students are examples of such self-sustaining patterns. The aim of perturbations, i.e., the intervention, is to durably change these self-sustaining patterns and replace them by new, more adequate patterns that, once they are established, should also be self-sustaining (Van Geert, 1994; 2003). From a dynamic systems point of view, changing these patterns of action and thinking of the teacher is quite similar to changing the patterns of action and thinking of the students, i.e., those can be indicated as teaching-learning processes.

In order to fully understand the effect of educational interventions on students' performance, insight is needed in the properties of these teaching-learning processes in individual teacher–student pairs. However, the progress of individual students as a result of an intervention is hardly reflected in effectiveness studies. This is because the effectiveness of interventions is usually studied using standard research practices. This methodological study aims to demonstrate how properties of a complex dynamic systems approach can help gain insight into change in teaching-learning processes due to educational interventions. This will be illustrated by examining a science education intervention, Video Feedback Coaching for teachers (VFCt), aimed at improving the quality of teachers' questions, and by doing so increasing students' scientific reasoning levels.

Standard Research on Educational Interventions

Assuming that the description given above provides a reasonably realistic picture of education as a complex dynamic system, we may ask ourselves what kind of picture teachers, parents and policymakers get from the standard research on education. Although probably few teachers will read the scientific journals on education science, the standard approach trickles down via various sources, such as via policymakers who have been trained in the standard practice of educational research, or the news media who report about scientific findings on education.

What the standard research practice implicitly or explicitly conveys to educators is, to begin with, the idea that influences of one variable onto another —such as motivation on school science performance— can be meaningfully separated from other influences and then in a sense stitched together again to provide a picture of individual educational processes.

Another idea that educators can get from the research is that effectiveness of an intervention (a curriculum, a teacher training program and so forth) resides in the intervention itself, i.e., that effectiveness is like an intrinsic causal force present in the intervention. In addition, the effectiveness of an intervention is something that is seen as applicable to particular kind of persons, i.e., to particular populations, such as the population of primary school teachers.

Another idea that the standard research practice in education conveys to educators is that knowledge and skills are internal properties of individuals, internal representations, internally represented schemes of action and so forth that are

transmitted from a teacher or a curriculum to an individual student. These internal skills or levels of knowledge can best be measured by validated, normed and relatively objective measurement instruments that express the internal skill or knowledge by means of a single number, i.e., a test score on a science test (Borman, Gamoran, & Bowdon, 2008 Penuel, Gallagher, & Moorthy, 2011; Şimşek & Kabapinar, 2010). Though, a more proximal measure, at the behavioral level, like the quality or complexity of the answers may be a better indicator of, for instance, a student's scientific reasoning level compared to a more distal measurement, like paper and pencil tests —as paper-and-pencil tests require other skills like reading as well (Van der Steen, Steenbeek, Van Dijk, & Van Geert, 2015). In addition, several studies report that interaction is essential to stimulate students' performance (Vygotsky, 1986). More specific, both Chin (2006) and Oliveira (2010) state that asking thought-provoking, student-centered, questions is a key element to stimulate students to reason with longer sentences and on higher levels of understanding.

Standard educational research also conveys the idea that what actually matters is the real or true skill, level of knowledge or ability, and that this real or true skill or ability can best be represented by averaging over individual fluctuations or individual variability (for more information see Rosmalen, Wenting, Roest, De Jonge, & Bos, 2012). The message is that these fluctuations or variability are in fact purely random variations around the true skill, level of knowledge or ability, and that they reflect purely accidental influences. For that reason, such fluctuations or variability within individuals are not intrinsically interesting, and should thus be averaged out. Preferably this is done by averaging over many individuals who, together, constitute a representative sample of the unit of analysis that really matters, namely the unit of populations characterized by a particular natural property, such as "typically developing students" or "dyslexia."

In this standard approach, there is of course room for interaction, for context, for individual variation, for change over time and so forth. These aspects are, however, clearly viewed from a perspective that is different from the perspective of complex dynamic systems. In the latter, they are like the primary givens, the starting point of theory formation and research (Fogel, 2011; Thelen, 1992; Van Geert, 2003), whereas in the more standard picture they are like secondary aspects, inferred from the primary aspects of research as discussed above.

How should educational research be transformed in such a way that it can convey to educators a picture of education that comes closer to the reality of education as a complex dynamic system? In the remainder of this chapter, we shall first discuss how properties of a complex dynamic systems approach can be applied to study the effect of educational interventions, such as the Video Feedback Coaching program for teachers. This approach will then be further illustrated by discussing an example of educational research, which uses properties from complex dynamic systems thinking in order to examine the effect of an intervention.

Intervention Assessments

In order to assess the effectiveness of such interventions several guidelines are frequently used. Veerman and van Yperen (2007), for instance, describe an often used classification scheme for assessing the effectiveness of youth care interventions as evidence-based practice. This scheme consists of four stages from potential effective interventions to efficacious interventions. An intervention is considered effective when the causality between the intervention and the outcome can be determined. Large-scale experimental research, multiple case-studies and norm related research are considered as ways to accomplish these causal relations.

Another way to establish the effectiveness of an intervention has been described by Boelhouwer (2013; as adapted from Lichtwarck-Aschoff, Van Geert, Bosma, & Kunnen, 2008). Boelhouwer proposes a taxonomy using four dimensions —which are grounded in the complex dynamic systems approach— to address the effectiveness of an intervention. Boelhouwer stresses the importance of using observational data and studying mutual causality. The four dimensions are:

1. *The static versus dynamic dimension* pertains to the dimension of analysis. Respectively, data are aggregated over many individuals versus data are displayed as a process over many time points. The *static dimension* can be used to analyze science performance as a combination of factors in a large sample. The effect of an intervention can, for instance, be assessed by focusing on the difference-score between pre measure and post measure, in which half of the participants receive an intervention while the other half does not (control group). The *dynamic dimension*, on the other hand, can be used to depict the process of change. Time series are used to depict how the changes emerge in and over time (Velicer, 2010).
2. *The micro versus macro time-scale* refers to the time-dimension. Respectively, a student's performance in real-time (i.e., the micro time-scale of seconds, minutes or hours (Lewis, 1995)) versus learning and development over several lessons or years (i.e., the macro time-scale of weeks, months or years (Lewis, 1995)). Analysis can be situated on different time scales at which the micro level is at the one end of the continuum and the macro level on the other end of that continuum. At the *micro level*, scientific reasoning skills can be captured in one specific situation, in which action sequences are studied. An example is a conversation during a science and technology lesson, consisting of one or several action–reaction sequences. At the *macro level* scientific reasoning skills can be captured over a longer period of time, for instance a series of science and technology lessons. The change in students' scientific reasoning skills due to the implementation of an intervention can also be interpreted as an example of a macro time-scale.
3. *The distinction between direct and indirect assessment* refers to the dimension of information sources, respectively the assessed person him or herself or a third-party assessor. A researcher can use several sources of information when evaluating an intervention program. One way is using *direct measures*, which

means information from those persons who actively participate in the intervention. In a professionalization trajectory for teachers, the teacher would be a direct source of information when (s)he is observing own behavior and reports about that, for instance by means of a questionnaire. *Indirect assessment* might refer to scientists who report about behavioral change.

4. *The distinction between short-term effects versus long-term effects* refers to the dimension of behavioral change due to —the effects of— an intervention. The short-term effects of an intervention can be seen as a change in observable behavior right after or eventually during the intervention lessons. The long-term effects refer to maintaining effects that are still observable a long time after the intervention, which can be visualized at follow-up or post-measurements (Boelhouwer, 2013; Steenbeek & Van Geert, 2015).

Using a Complexity Approach to Map Change: How to Apply the Properties

The complex dynamic systems approach offers tools to focus on properties of development and learning as dynamic processes (Steenbeek & Van Geert, 2013; Van Geert, 1994), which lie beneath the aforementioned dimensions. Using this approach is a way to study how learning occurs in interaction with the material and social context by focusing on those processes during real-time and frequent observations, i.e., during actual lessons (Granott & Parziale, 2002; Van Geert, 1994; Van Geert & Fischer, 2009). In order to understand the dynamics of a complex system, such as a teacher's behavior in the context of a group of developing students, the assessment should also focus on the dynamic character of learning, i.e., how a student's performance emerges in interaction with the context (see Steenbeek & Van Geert, 2013; Wetzels, Steenbeek, & Van Geert, in press). Observational methods, i.e., video recordings, are considered essential to be able to capture the developments on these real-time (micro) timescales and to preserve the complexity of the process of learning. Several properties of learning —such as change, nonlinearity, iteration and self-organization, variability, and the transactional nature of learning— as a result of an intervention must accordingly be taken into account. Mapping these properties is important to explain average group-based findings and provide insight into the underlying processes of learning and subsequent performance of individual students (Van Geert, 2004) and the quality of a science education intervention (Wetzels et al., 2015). The relevance of a complex dynamic systems approach, for intervention studies, demonstrates itself in offering possibilities for answering different research questions.

In the next section we will discuss three important properties of a complex dynamic system for the context of learning. This is a background for understanding the need for a process-based methodology. For this reason, we will describe how

underlying properties of Boelhouwer's (2013) dimensions can be integrated in educational intervention studies as an essential addition to group-based analyses.

The role of time in change has a prominent role in Boelhouwer's taxonomy: in the time-dimension (micro versus macro) as well as the behavioral change-dimension (short-term versus long-term intervention effects). Velicer (2010) states that a time series analysis can help to understand the underlying naturalistic process and patterns of change over time, or to evaluate the effects of an intervention. For instance, time provides valuable information about the dependency between all measurements. As Steenbeek and Van Geert (2005) state, behavior of the student —which can be as small as an utterance— at a certain point in time affects the subsequent activity of the teacher —also known as *iteration*.[1] Since changes in the micro-timescale —short-term effects— are intertwined with long-term effects, analyzing student's actions during real-time interactions might be helpful in understanding change (Steenbeek, Jansen, & Van Geert, 2012).

As an illustration, let us return to a science class in an upper grade elementary classroom. The teacher's questions influence the reactions of the students in the form of answers, signs of interest or of avoidance, which on their turn influence the subsequent questions and reactions of the teacher following the reactions of the students. Students hear other students giving an answer, or see them performing particular activities, and this influences their own potential answers to questions asked by the teacher. The effect of the interactions takes place on various, nested timescales (e.g., Van der Steen, Steenbeek, & Van Geert, 2012). There is, for instance, the short-term time scale of a particular science class, which involves the dynamics on the level of activities, solving problems and formulating explanations. There is also the long-term timescale of changes in the nature of the answers or the probabilities of high-level reasoning that develops as a consequence of the short-term interactions. As is typical of a complex dynamic system, events on these various timescales affect one another, that is to say there is mutual influence and reciprocal causality (Steenbeek et al., 2012). Another example is the short-term timescale of asking a particular kind of questions by the teacher and the long-term timescale of eventual changes in the nature of the questions asked by the teacher, for instance as a consequence of an intervention aimed at teacher professionalization (e.g., Wetzels et al., 2015). A class of students with their teacher tend to evolve towards particular, class-specific patterns of activity, that is to say towards a typical pattern of asking questions, giving assignments, giving answers, showing interest or

[1] Dynamic processes are *iterative* in nature. Iteration refers to "a procedure that operates on an input that is in fact its preceding output" (Van Geert, 1997). This means that over time, the teaching-learning process (the current state) is a product of the previous state, and serves as input for the next state. Teacher and student mutually influence each other over time; the current action of the teacher influences the next (re)action of the student, which influences the next (re)action of the teacher, and so on.

boredom, and many other properties. These patterns form some sort of complex *attractor state*[2] (e.g., Steenbeek & Van Geert, 2005) that is typical of the teacher-classroom system in question. These attractor patterns are in a sense self-sustaining, for instance the nature of the questions habitually asked by the teacher influences the nature of the answers habitually given by the students, and these answers are likely to sustain the nature of the questions asked by the teacher. In addition, the attractor patterns, i.e., few variability is visible in the teacher–student interaction patterns, are relatively resistant to change.

A focus on *variability*[3] provides information about interindividual variability and intraindividual variability. Bassano and Van Geert (2007) state that "variability is informative on the nature of developmental change". The dynamic dimension in Boelhouwer's taxonomy (2013) allows further for possibilities to map *inter-individual variability*[4], variability among students, teachers, or groups. This might be done to compare several individual teachers to find out whether one teacher's intervention trajectory is more effective compared to a similar intervention trajectory of another teacher. Questions might focus on whether the pathways of all students are equivalent, i.e., did they develop in similar ways? A change in student's science performance might be found in trajectories in which a teacher seems capable of adjusting his/her questions to a student's level of functioning and thinking, while the less effective trajectories remain in a fixed pattern of non-differentiating interactions (Ensing, Van der Aalsvoet, Van Geert, & Voet, 2014). Variability at the micro level (adjusting to the level of students) might, in this case, be an important element accounting for the variability between the teachers. Interindividual variability can provide important information about underlying dynamics of (less) effective intervention trajectories. Each trajectory —either an intervention or another developmental trajectory— takes the form of a dynamic pathway, constructed as real-time iterative processes, which emerges through interaction with the context (Fischer & Bidell, 2006). As each student starts an

[2] An *attractor state* is a temporarily stable state that recurs over time: "the state to which systems are attracted, that is, towards which they spontaneously evolve as a consequence of the underlying dynamic principles that govern their behavior" (Van Geert, 2003). For instance, in a classroom a teacher may routinely ask knowledge-based questions. This mode of interaction becomes a self-sustaining comfortable state for both teachers and students, making this type of questioning and students' reactions an attractor state for this particular classroom. If the teacher, under influence of the intervention, begins to change her questioning strategies towards open-ended questions, the students might first resist. However, if the teacher persists in using these open-ended questions and the students start to engage in critical thinking, the classroom system (teacher–student interaction) might change permanently over time—resulting in a new attractor state.

[3] *Variability* is defined as the "coexistence of many different patterns of development" (Van Geert, 1998). Two types of variability can be distinguished: 1. *Interindividual variability*: differences in the behavior between—groups of—individuals at some point in time. 2. *Intraindividual variability*: Van Geert and Van Dijk (2002) have defined intraindividual variability as "differences in the behavior within the same individuals, at different points in time" (p. 341).

[4] Note that 'individual' does not necessarily refer to a single person. It refers to the level at which a particular process actually occurs, which can be an individual person, but also a classroom.

intervention at their own level and masters science and technology to the best of his/her capabilities, each trajectory is unique and should be analyzed as such to provide insight in the variability.

Intraindividual variability is defined as "differences in the behavior within the same individual, at different points in time" (Van Geert & Van Dijk, 2002). By looking at multiple measures of individuals it is possible to see how the change and development proceeds (e.g., Van der Steen, Steenbeek, Van Dijk, & Van Geert, 2014). Van der Steen, et al. (2014), for instance, showed that a student's performance changed over several science activities. By focusing on intraindividual variability, a change in interaction between a student and a researcher was found. At the start of the learning trajectory, the teacher took initiative by asking thought-provoking questions during inquiry activities (state 1); the student followed the level of the teacher. At the third lesson, a change in interaction pattern was found, in that the student took initiative (state 2) and seemed to have initialized the process of inquiry. In between these two states, some form of "chaos", in this case increased variability, was found in which the researcher and student did not seem to adapt to each other as well as before (state 1) and after (state 2). Transitions from one state to another are often accompanied by qualitative indicators, but also by increased variability or critical slowing down of variability (e.g., Bassano & Van Geert, 2007).

The surplus value of focusing on variability is that it yields information about the differences in underlying characteristics leading to differences between lessons or participants. Specifically, this might show whether there are behavioral characteristics accounting for why a trajectory seems to yield more positive change for one subgroup than another or how one state changes into another (concerning development - Lichtwarck-Aschoff, Van Geert, Bosma, & Kunnen, 2008; education - Steenbeek et al., 2012; sports - Den Hartigh, Gernigon, Van Yperen, Marin, & Van Geert, 2014).

The *transactional nature* provides insight into how the learning gains of students can be understood, i.e., how is performance (co-) constructed during actual lessons, why is the intervention for some classrooms or students more effective than others? Learning can be seen as a dynamic and distributed, transactional process (Steenbeek, Van Geert, & Van Dijk, 2011). Students do often not come to a conclusion spontaneously. Teacher support is essential to reach a higher level of performance (Van de Pol, Volman, & Beishuizen, 2011). Teaching and learning are dynamic processes that are constantly adapting to changing needs and opportunities. It is therefore important to focus on the dynamics of reaching a performance by studying interactions, i.e., what the teacher's contribution is in students' performance. The unit of analysis ought to be the dyad of a teacher and the students, and not the individual student on its own. Knowing more about how teachers stimulate students toward higher levels of science performance might provide valuable information about how to optimize inquiry-based learning situations (Van der Steen et al., 2014).

Note that although the three properties describe distinct mechanisms, during the process of learning, they all work simultaneously. Boelhouwer's (2013) observational dimensions might be seen as different levels of analyses and can show

increasingly detailed information about how well the averaged findings (static) represent the variability in individual trajectories (dynamic) in (micro) and over time (macro). For the purpose of this article, the properties are presented in such a way that the surplus value compared to the classical approach is stressed (Table 11.1). However, we do not intend to give the impression that this classification is *the* ultimate way to study interventions. The principles of variability can, for instance, be very well applied at the micro level to find change points in the transactional nature of learning trajectories over several lessons (e.g., Steenbeek et al., 2012).

Present Study

In this study, we aim to demonstrate the contribution of a complex dynamic systems approach when assessing the effectiveness of a science education intervention. We illustrate this by presenting data from the Curious Minds Video Feedback Coaching program for teachers (Van Vondel, Steenbeek, Van Dijk, & Van Geert, 2015). By doing so, we intent to provide a more thorough and multifaceted view of the process of studying the effectiveness of an intervention, compared to standard evaluations. By starting with the more classical group-based analysis, we aim to demonstrate that each remaining analysis —increasingly more process-based— can provide more understanding of the effectiveness. Hence, information about students' performance (static and dynamic) and the development of students' scientific reasoning skills during one lesson (micro) and over several lessons (macro) will be presented. In addition, the role of the teacher in this process (micro-dynamic) can be shown during the intervention (short-term effects) and a few weeks after the intervention (long-term effects).

Method

Rationale for the Teaching Intervention

The Video Feedback Coaching program for upper grade teachers is a professionalization trajectory designed to support teachers in improving the quality of science education lessons in their classroom. More specifically, this pedagogical-didactic intervention was developed to stimulate change in teacher–student interactions, i.e., changing the discourse from mostly teacher-centered into a more stimulating student-centered discourse (Wetzels et al., in press). By doing so, teachers enhance the quality of students' scientific reasoning skills by establishing a series of inspiring teachable science moments (Bentley, 1995; Hyun & Marshall, 2003). The way teachers interact with students was regarded as a key to quality of the

Table 11.1 Combination of complexity properties and dimensions as formulated in Boelhouwer's

			Dimensions			
	Goal	(possible) Research question	Information source	Analysis	Time	Behavioral change
Pre versus post-measure	Generalization Insight into whether there is an effect or not	What is the group-based effect of the intervention?	Indirect group	Static aggregated	Macro	Long term effects
Role of time in change	Map development and/or change Help to understand the underlying naturalistic process and patterns of change over time, important to evaluate the effects of an intervention	How can we characterize short-term and long-term change in students' scientific reasoning on group level during the intervention trajectory?	Indirect group	Dynamic	Macro	Short- and long-term effects
(Intraindividual) variability	Map temporal change Information about the quality of interventions	How does this classroom level change over the time covered by the intervention?	Indirect classroom	Dynamic	Macro	Short- and long-term effects
Transactional nature	Map co-construction Insight into how the learning gains of students can be understood	How is scientific reasoning co-constructed and how does this co-construction process change over time under influence of the intervention?	Indirect individual dyad	Process (dynamic)	Micro	Short-term effects

science lessons. The intervention contained the following evidence-based key elements: (1) improving teachers' knowledge about teaching science and scientific skills, (2) establishing behavioral change by improving teachers' instructional skills by means of (a) VFCt and (b) articulating personal learning goals.

The first element was reflected in an interactive educational session about knowledge of teaching science and scientific skills for participating teachers. Osborne (2014) defined these skills as knowledge about the process of science —including knowledge about the empirical cycle— and the skills needed for performing an actual scientific inquiry —such as higher order thinking skills. During this educational session information was provided and the features important for science learning were discussed: the use of the empirical cycle (De Groot, 1994), use of thought-provoking questions (Chin, 2006; Oliveira, 2010), scaffolding (Van de Pol et al., 2011), and science and technology-education in general (Gibson & Chase, 2002). According to Lehmann and Gruber (2006) expertise can best be acquired through case-based learning, including authentic cases which are embedded in naturalistic contexts. Therefore, several best-practice video fragments of teacher–student interactions during science lessons were shown to illustrate the transactional nature of performance; i.e., the importance and effect of high quality interactions during science and technology-activities.

The second element referred to the aim to establish —durable— behavioral change. A promising method for implementing evidence-based instructional strategies, i.e., establishing behavioral change is providing feedback on real-time behavior (Noell et al., 2005; Reinke, Sprick, & Knight, 2009). Teachers instructional quality can be greatly increased by offering video feedback on own classroom behaviors (see also Mortenson & Witt, 1998; Seidel, Stürmer, Blomberg, Kobarg, & Schwindt, 2011; Wetzels et al., 2015). As a rule, the effect of feedback is best when a 3/1 ratio is used (Fredrickson, 2015), i.e., three positive fragments were discussed and one fragment which could be improved. In order to stimulate teachers to fully understand the behavioral patterns and consequences of those interactions for students' performance, the coaching focused on the transactional nature of learning by reflecting on teacher's own specific behaviors and interactions at the micro-timescale and was conducted immediately after each lesson, as immediate feedback is most beneficial for learning (Fukkink, Trienekens, & Kramer, 2011). Note that aside from this practical application, these videotapes were used as the primary source to evaluate the effectiveness of the intervention.

In addition, goal setting at the beginning of a coaching trajectory is an effective way to achieve results (Hock, Schumaker, & Deschler, 1995), i.e., behavioral change, as they ensure feelings of autonomy (Pintrich, 2000). By formulating learning goals that reflect teacher's personal professionalization trajectory, teacher's feelings of autonomy were respected and teachers were provided with opportunities to monitor and control their motivation and behavior. Another way to ensure teacher's feelings of autonomy and thus to create more responsibility for their own learning process, was by encouraging them to prepare science and technology-lessons to his or her own liking (Table 11.2). Teachers were allowed to choose a topic and an instructional method (for instance experiments or a design

Table 11.2 Type and topic of lessons as provided by each teacher

	Pre-measure	Lesson 1	Lesson 2	Lesson 3	Lesson 4	Post-measure
Classroom 1	Experiments: air pressure	Classical experiment: air pressure	Classical experiment: air pressure	Classical experiment: air pressure	–	Classical experiment: air pressure
Classroom 2	Experiments: air pressure	Experiments: surface tension	Design: planetarium	Drawing: rainbow	Experiments: gravity	Experiments: air pressure
Classroom 3	Experiments: air pressure	Experiments: air pressure	Experiments: gravity	Experiments: gravity	Experiments: balance	Experiments: air pressure
Classroom 4	Experiments: air pressure	Classical experiment: air pressure	Classical experiment: air pressure	Classical experiment: air pressure	Classical follow-up discussion: air pressure	Classical experiment: air pressure
Classroom 5	Design: barometer	Experiment: air pressure	Classical experiment: air pressure	Experiment: water	–	Classical experiment: air pressure
Classroom 6	Experiment: air pressure	Laptop: satellite	Design: balloon rocket	Experiment: blending	Design: "Techniektoren"[a]	Experiments: air pressure

[a]The "Technique Towers" are lockers shaped as towers with 80 lesson-boxes inside—10 lessons for each year of Dutch elementary education. Each lesson box is focused on a specific aspect within a domain of technology, for instance construction or making soap. Each box has a step-by-step manual which can be used by a small group of students, without a teacher. This manual guides them through the activity that is in the box

assignment) suiting their own and students' interest. The table shows that the first lesson of classroom 6 mainly focused on experiments and the topic was air pressure. The main focus of the next lesson was on using laptops to search for information about satellites.

Participants

Six upper grade teachers (two men and four women) and their students (M_{age}: 11.2, 9–12 year olds) from the North of the Netherlands participated in the study in school year 2013/2014. Their teaching-experience ranged from 6 to 18 years in regular elementary education. The average classroom consisted of 28 students (49 % girls, 51 % boys).

Procedure and Materials

Six science and technology lessons and an educational session were conducted in a period of 3 months: one pre-measure, four lessons immediately followed by a VFC session led by a trained coach (first author) and one post-measure, on average 4.5 weeks, after the end of the VFCt.

Although the intervention was intended as adaptive support and was highly idiosyncratic, some standardization was implemented during data collection. That is, the same coach provided identical information during the introductory session, videotaped all lessons, and was responsible for the guided reflection after each lesson. In addition, teachers were asked to use the following guidelines: provide six lessons using a fixed format: introduction (plenary introduction), middle part (students work on their own or in groups), and end (plenary discussion). Furthermore, they were asked to teach lessons about the "earth and space" system —such as weather, air pressure, gravity, or the positions of the moon. Lastly, the teachers were instructed to focus on air pressure and aim at learning students about high and low pressure during the pre-measurement and post-measurement.

Data Analysis

Ten minutes of the middle part of the lessons were coded, because in this part a relatively larger amount of rich, interactive interaction was present. For further data analysis, the classroom of students as a whole was taken as the unit of analysis, which means that the individual case is always consisting of a group of individuals. However, in contrast with the classical group approach of looking at the performance of independent individuals, which most studies use to calculate averages,

this group is conceived of as a collection of interdependent individuals interacting with each other. In line with that, the previous utterance of the teacher or a fellow student was taken into account when scoring students' level of complexity.

Students' scientific reasoning skills were measured by quantifying verbal utterances, using a scale based on skill theory (Meindertsma, Van Dijk, Steenbeek, & Van Geert, 2012; Parziale & Fischer, 1998; Van der Steen, Steenbeek, Wielinski, & Van Geert, 2012). The dynamic skill theory (Fischer, 1980) is a cognitive developmental theory focusing on how skills —which are considered complex and variable— are constructed in specific domains. These skills can be captured by focusing on those skills as they emerge in interaction with the context. This scale has proven useful for task-independent measures in the analysis of student's scientific explanations. Student utterances were scored on complexity using a 10-point scale, divided in three tiers (sensory-motor, representations, and abstractions). The first tier (level 1–3) consists of sensorimotor observations and explanations, which mean simple observable connections are given. Level 1 means the least complex utterance, a single sensorimotor aspect (e.g., an expression of what they see; the student says: "It [the balloon] is white"). At level 2, the sensorimotor mapping level, the student is able to combine to single sensorimotor aspect into one mapping (e.g., the student says: "It is white and that one is yellow."). The second tier (level 4–6) comprises representational predictions and explanations, which means that students use higher order thinking skills to go beyond simple perception-action couplings. The student understands that an object has a specific characteristic, outside the present situation. (S)he can, for instance, make a prediction about what is going to happen when you put salt into a water/oil fluid —without directly seeing it. The third tier (level 7–9) constitutes abstract explanations; students are capable of generalizing ideas about the object outside specific situations. A student might for instance explain that "the molecules in the water are strongly drawn towards each other... probably leading to surface tensions... the water and oil cannot blend because of that" or "the density of the water is higher compared to the density of the oil, the fluid with lower density floats". Level 10 could be scored when students expressed understanding about global laws and principles (e.g., the abstract principles of thermodynamics can be applied to the situation at hand). Ten to twelve-year olds are expected to be "capable" of reaching the seventh level of understanding (Fischer & Bidell, 2006). They could express abstract thinking skills (e.g., relate abstract concepts to the situation at hand, as showing in the following utterance "the air pressure pushes the paper towards the table").

Coding was done by means of the program "Mediacoder"[5] (Bos & Steenbeek, 2009). To establish the interobserver reliability for the application of the coding scheme, the interobserver agreement was determined in advance by the first author

[5] Mediacoder can be obtained free of charge by sending an e-mail to one of the developers: h.w.steenbeek@rug.nl or j.bos@rug.nl. Mediacoder is a simple application for coding behaviors within media files. A media code is a moment in time in a recorded video when a particular event occurs. The meaning of the coding is determined by the specified character (which the user can choose him/herself). The point in time is determined by the time within the recorded media. Each media

and an independent coder. With an agreement ranging from 79 to 83 %, Cohen's kappa of .76, the interobserver agreement was considered substantial.

Excel was used for descriptive analysis and to display patterns in the data. As the collected data consisted of a small group of participants, dependency between variables, and multiple measures, a nonparametric test was used to test differences in students' scientific reasoning level over several lessons. This random permutation test was used to test the empirical results in relation to a statistically simulated baseline of random patterns, using Poptools (Hood, 2004). This means that the nonparametric test statistically simulated the null hypothesis that the probability of the relationship or property was based on chance alone. For instance, the scientific reasoning level data were randomly shuffled (values were randomly drawn from the data without replacement), and the same average and difference score was calculated for the statistical simulation of the null hypothesis. This random shuffling, i.e., data generated on the basis of the null hypothesis model that there was no effect of the intervention, was permutated 1000 times in order to calculate whether the empirically found difference between pre- and post-measure could be expected to occur on the basis of chance. When the finding was smaller than 0.05, the test statistic was considered significant. This means that when we speak about significantly different, we mean a considerable difference that has applied meaning (for instance a difference that is big enough, one complexity level, for the teacher to be observed in the real world). A significance score between 0.05 and 0.1 is considered as a trend, i.e., non-randomness (see for a discussion about cut-off scores of *p*-values and the use of confidence intervals: Kline, 2004; Lambdin, 2012; Cumming, 2014).

Pre-measure versus Post-measure

All task-related student utterances were coded on complexity. Subsequently, we calculated the average complexity level of all students over all classes at pre-measure and post-measure, and computed the difference between the two. In addition, as significance scores are not directly linked to practical significance (Sullivan & Feinn, 2012) the effect size was calculated using Cohen's D. Following Sullivan and Feinn, an effect size of 0.2 is considered small, 0.5 medium, 0.8 large, and 1.3 or higher very large.

The Role of Time in Change

The long-term effects were operationalized as the effects that were still observable, 4.5 weeks, after the intervention. These were assessed by comparing students' scientific reasoning level at the intervention-lessons with students' scientific

code can be supplemented with an explanation. After coding the file can be exported to excel or SPSS for further analyses.

reasoning at the post-measure. Therefore, we calculated the average complexity score of each lesson. The same was done for the statistical simulation of the null hypothesis. Short-term effects were assessed by focusing on scores during the intervention.

Variability

Again, all students in the classroom were taken as our unit of analysis, and focused on the classroom performance level. While doing so, the focus was on variability in the sense of differences between the various classrooms (interindividual variability) and of differences over time within classrooms (intraindividual variability). The variability of each classroom was computed and compared with the variability between lessons of that classroom. The same analysis was done on the group level, in that the variability was computed of all classrooms and the variability of each classroom was compared with the overall —averaged— variability. This analysis can be the basis to find intraindividual variability which might show the properties of effective and less effective trajectories. In order to actually study the process, you must study the process on the individual case level. Second, in an attempt to generalize, or more precisely to find similarities between individual cases, clustering techniques may be used (e.g., clustering of students working on science activities; Van der Steen et al., submitted). As an illustration a simple example of looking for groups of cases, of which the averages are clearly different, will be presented. The quantitative findings were supplemented with qualitative findings, derived from video fragments, to show possible explanations for variability between and within classrooms (mixed method; Johnson, Onwuegbuzie, & Turner, 2007). Significant differences were used as a starting point for examining the data in a qualitative manner.

Transactional Nature of Learning

In order to be able to make a comparison with the first, group-based analysis, the focus of this representative case was again on the pre-measures and post-measures. Variables which were assessed (over time) concerned task-related utterances: the number and types of questions asked by the teacher, the complexity of student utterances, and the occurrence of coherent “action–reaction chains” in teacher–student interaction. Therefore, for the teacher variable the utterances were coded on an ordinal scale of “level of stimulation” (based on the “openness-scale” of Meindertsma, Van Dijk, Steenbeek, & Van Geert, 2014); i.e., utterances intended to evoke students’ (higher order) scientific reasoning skills. The scale ranged from giving instructions, providing information, asking a knowledge-based question, asking a thought-provoking question, posing encouragements, to posing a task-related follow-up. Giving an instruction is considered as least stimulating, i.e., the smallest possible chance of evoking a high level of reasoning as an answer. With a Cohen’s kappa of .72 the interobserver agreement was considered substantial. First,

the interactional space, i.e., the amount of utterances, covered by the teacher and students was computed to gain insight into the general distributions of turns during the lesson. Note that the non-task related utterances are removed from this graph. Next, a graph showing the temporal sequence of the interaction is displayed (with the program Excel), as an alternative to the state space grid method (Hollenstein & Lewis, 2006). Both the graph and a state space grid use two axes to display the interaction between variables. A state space grid is a useful way to depict attractor states. However, for the purpose of answering the research question about how scientific understanding is co-constructed an excel graph is, in this particular case, a more accessible application. Lastly, a transition diagram (e.g., Ensing et al., 2014; Steenbeek et al., 2012) was used to study the micro dynamics of the transaction between students (as a class) and the teacher. Transition diagrams were made to reveal pattern characteristics, which provide insight into the number and types of questions asked and potentially how the difference between pre- and post-measure can be explained. These diagrams show the succession of variables. The observed differences between the pre- and post-measure regarding the percentages were statistically tested based on the null hypothesis that the observed differences were accidental. For the transition diagrams the follow-ups were summarized in non-stimulating reactions —instructions, providing information— and stimulating reactions —thought-provoking questions and comments and encouragements.

Results

Pre-measure versus Post-measure: Static-Macro Dimension

In order to answer the research question on whether there is an effect of the VFCt on students' performance, the observational data of the pre- and post-measure is aggregated over all classrooms. Note that the pre-measure and post-measure had the same teaching goal in all groups, i.e., teaching students about high and low (air) pressure. The scores during these lessons can therefore be compared validly.

Students performed on average better during the post-measure, $M = 4$, compared to the pre-measure, $M = 3.25$ ($p < 0.05$; Cohens $d = 1.6$, very large). Results show an expected intervention effect, i.e., students' science performance increased. This static macro dimension is the standard answer to questions about effectiveness of an intervention; most researchers are confining themselves to this single static macro evaluation. However, more insight can easily be gained by knowing how these average classroom complexity levels are constructed. In this particular case, the lower levels of scientific reasoning (1, 2, 3) are, for instance, more apparent during the pre-measure (PreM = 52; PostM = 25), while the higher levels (5, 6, 7) of scientific reasoning (PreM = 17; PostM = 36) are manifested more during the post measure (resp. $p < 0.05$ and $p < 0.01$). Looking at all measurements provides more information about the question what happens during the intervention-lessons.

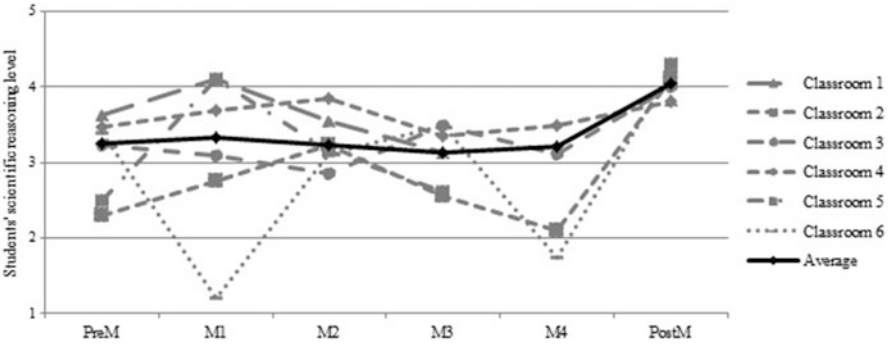

Fig. 11.1 Dynamic-macro scores of students' scientific reasoning skills of all classrooms during all measurements

Time: Short- and Long-Term Effects

In order to answer the question about development; how can we characterize students' scientific reasoning on the group level during the intervention trajectory, the solid black-diamonds line in Fig. 11.1 represents the average score of students' scientific reasoning level over all classrooms over time.

The solid line in Fig. 11.1 depicts that students display higher levels of scientific reasoning at the post-measure compared to the other measurements (preM = M1 = M2 = M3 = M4 < postM, $p < 0.01$). We thus see a long-term effect for this variable and the level of scientific reasoning seems rather stable on group level from the pre-measure to the lessons during the intervention.

This is already one step forward in comparison to the static macro comparison of the pre- and posttest. However, since the black line represents the average of the levels for all the classes, it is still the representation of a pseudo process (as a sequence of averages over independent cases it is not a real process). Based on this notion of a pseudo process, in order to actually see the process of change, analysis should focus at the process on the individual level, which in this case is the classroom level. Note that this is, in turn, a pseudo-process for the individual trajectories.

Variability: Dynamic-Macro Dimension

Next, there is a need to know the performance level of each classroom and how this changes (dynamic) over time (macro) under influence of the VFCt. Figure 11.1 depicts considerable variation in the level of scientific reasoning between classrooms (dashed lines), but also within a classroom over time.

With regard to interindividual variability: In Fig. 11.1, all observations over the six clasrooms in the post-measurement case are very close to one another, whereas

almost all the preceding measurements show quite considerable variation between individual classrooms. This shows, for instance, that during post-measure the average complexity level of students' level of scientific reasoning of all classrooms is closer to each other compared to the pre-measure ($p = 0.1$). Furthermore, quite considerable differences were found in the amount of task-related utterances among classes. For instance, classroom 6's first scientific reasoning level is based on five task-related utterances ranging from complexity level 1 to 4, while classroom 1's level is based upon 54 task-related utterances ranging from complexity level 1 to 7. In addition, as an illustration of the clustering of individual cases: two subgroups were found in the level of variability ($M1_{variability} = 0.4$ and $M2_{variability} = 1.3$, $p < 0.01$). Classroom 1, 3, and 4 showed a rather stable level of scientific reasoning level over the lessons ($M_{variability} = 0.4$), while classroom 2, 5, and 6 showed considerable variability ($M_{variability} = 1.3$).

With regard to intraindividual variability: Intraindividual variability is visible in all classrooms (see Fig. 11.1, dashed lines), but most clearly in classrooms 2, 5, and 6 (note that this is one of the two subgroups mentioned above). When we zoom in at the development of classroom 6, the difference between the first and second lesson in students' scientific reasoning level is 1.91 complexity level. Measurement 1 ($p < 0.01$) and measurement 4 ($p < 0.01$) are different from the other lessons in that the average scientific reasoning level is lower. During both lessons only a handful task-related utterances could be scored, and 75–80 % of those utterances were on the lowest complexity level.

Looking back, these results may be explained by the content of lesson 1 and 4 (Table 11.2—method section). In both cases, the students were not allowed to experiment and the material was less provoking (note that the same variation in lessons applies for classroom 2 and 5). This suggests that the type of lesson and material used influences the —amount of— emergent complexity level of students' utterances.

Transactional Nature of Learning: Micro-dynamic and Long-Term Effects

Due to the labor-intensive nature of the observations, the following illustrations focus on one representative case; one teacher and her students. Classroom 3 could be used as a representative case in that preliminary analyses of teacher behavior showed that the behavior of the teacher represented the general interactional patterns in the classrooms best —i.e., starting the intervention by predominantly using instruction towards a more thought-provoking teaching style at the end of the intervention— the teacher neatly followed the guidelines, students' average age closely resembled the average age of all participating students, and all measures were available of this classroom.

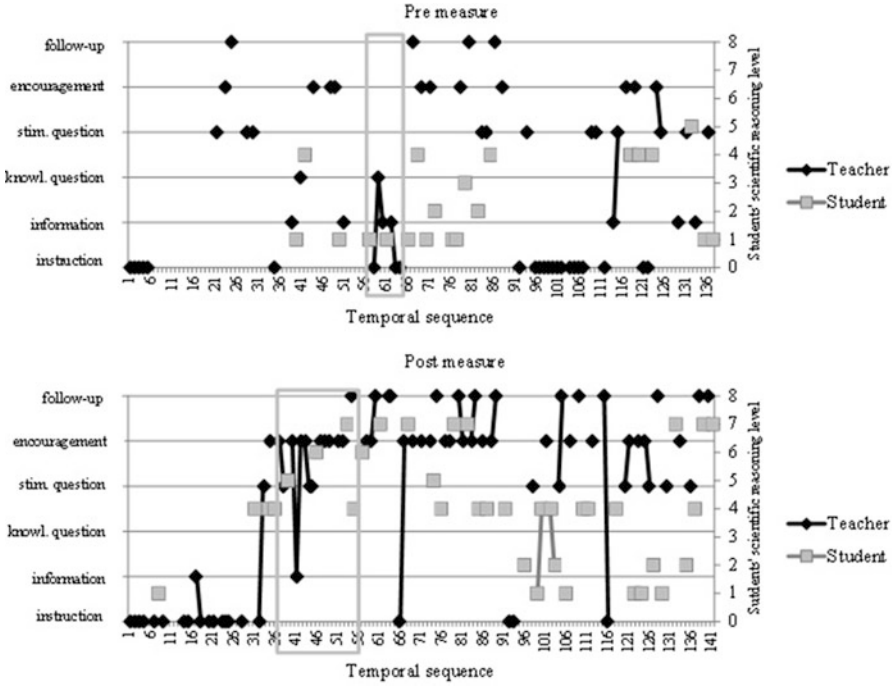

Fig. 11.2 Dynamic-micro scores during pre- (*top*) and post-measure (*bottom*) for classroom 3 *Note*: the teacher (*left*) axis depicts an ordinal scale from less stimulating to more stimulating utterances to provoke scientific reasoning skills: 0 = instruction, 1 = providing information, 2 = knowledge question, 3 = thought-provoking question, 4 = encouragement, 5 = follow-up. The student (*right*) axis depicts the ordinal complexity scale based on skill theory. The *grey boxes* are illustrated in the text

Figure 11.2 depicts the quantified interaction during 10 minutes of the middle part of the pre- and post-measure of classroom 3. The figure depicts different interaction patterns during pre-measure and post-measure. During post-measure there is in general much more interaction, mainly at the higher (more stimulating and complex) side of the graph. This type of display is a way to represent the nature of the process of interaction between the teacher and the students. On the *x*-axis the temporal sequence of the interaction is displayed. Each number represents an utterance of either the teacher or the student. On the left *y*-axis the task-related teacher utterances (diamonds) are categorized according to the degree of stimulation, while on the right *y*-axis the complexity level of task-related student utterances (squares) are depicted. Blank spaces represent a non-task related utterance. For purposes of illustration and as a guide how to read the graph, part of a literal translated transcript of an experiment "blow a paper wad in a bottle" will be described. Starting from utterance 57 (the grey square in Fig. 11.2, on the top): the teacher starts with a knowledge-based question: "*I think... What's in there?*"

followed by self-iterated information giving "*There is still moisture in it*." Next the student answers by formulating what he sees: "*Yes, it is red*." The teacher continues with providing information "*And then the paper sticks, that's a shame*." She offers a possibility for why the moisture has an effect on the outcome "*This bottle is dry...*" and offers a new bottle with the instruction to retry the experiment: "*Try this [dry] one*."

The level of stimulation: Figure 11.2, on the top, depicts that the teacher occupies most interactional space (75 %) during the lesson, more specifically most of her utterances are on the lowest stimulation level, namely to instruct students (41 % of her utterances). The transcript described above is an example of that type of interaction. During the pre-measure most of the utterances were teacher-centered (56 %), i.e., focusing on what students need to do and on knowledge acquisition by instructing, providing information, and asking knowledge-based questions, while 44 % were student centered utterances, i.e., stimulating utterances focusing on students thinking process—thought-provoking questions, encouragement, and substantive follow-up.

In contrast, although the teacher still occupies most of the interactional space (67 %) during post-measure, we can now see reciprocity between teacher utterances and student utterances which seem to emerge in higher levels of complexity (Fig. 11.2, bottom). Compare for this the upper side (on the teacher axis stimulating question, encouragement and follow-up) of the pre-measure graph with the upper side of the post-measure graph. During post-measure there is much more interaction at the higher (more stimulating and complex) side of the graph. The teacher asks more questions, poses more encouragements and students reason on higher levels of complexity (4, 5, 6, and 7). In addition, compared to the pre-measure a reversed pattern was found in teacher style, meaning that 29 % was teacher-centered (least stimulating) and 71 % consisted of stimulating utterances during post-measure. Table 11.3 describes an interaction during the post-measure showing how this was seen during the activity. Here, the teacher starts the interaction with a thought-provoking question, followed by a student answer that shows understanding of the experiment. The teacher continues with encouragements and rephrases student answers.

To conclude, by comparing the pre-measure and post-measure, the quantitative data shows an emerging pattern in which the teacher uses higher levels of stimulation during post-measure. The teacher asks more stimulating questions or poses encouragements to reason further (compared to preM; $p < 0.05$), students answer more often (preM = 20; postM = 36) and on a higher level of complexity ($p < 0.01$).

Action—reaction sequences: Figure 11.3 shows transition diagrams of both lessons. Both the type and number of teacher and student utterances change. During the pre-measure, students answer a teacher initiation question in only 31 % of the cases and the teacher answers her own question or continues herself in 15 % of the utterances. A student answer is in 20 % of the cases followed by a non-stimulating teacher response (like providing information or instruction) and in 44 % of the cases by a stimulating follow-up (encouragement, question, or an utterance to encourage

Table 11.3 Literal translated transcript of the experiment: "candle and lemonade," starting from utterances 38 (grey square) and further during the post-measure

Teacher	Student(s)	Comment
What do you think [will happen] [***]?		Thought-provoking initiation question
	When you put the glass over [the candle]... the water comes up and... because of the water the candle goes out	Student is capable of making a representation about what he expects to happen
Ok... hmm...		Encouragements (without directing to the "right" answer)
Ok, you think the candle extinguishes because of the water.		Rephrasing student's answer
Who has another idea?		Invite other students to formulate a hypothesis
	When you put the glass... the fire causes vapor... when that comes down the candle stops burning	Student is capable of formulating a representation in which insight into a natural phenomenon is represented
Hmm... Basically you make rain...		Rephrasing student's answer — providing information about how it could compare to daily life situations
What do you think [***]?		Invite another student to formulate a hypothesis
	I think there will be no more oxygen	Student is capable of formulating a hypothesis using abstract language
No more oxygen... Where?		Teacher uses a follow-up question to make the student elaborate on her answer

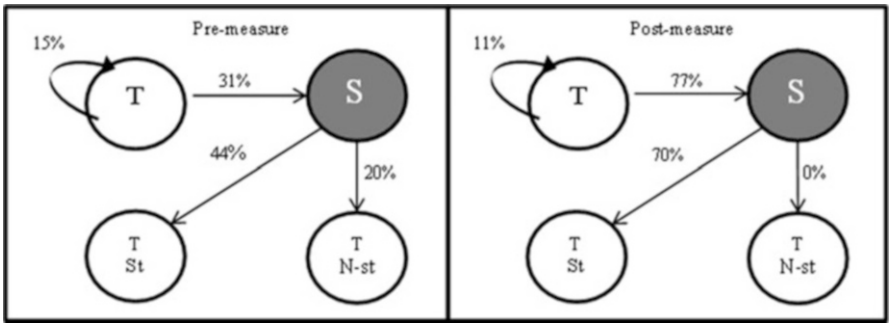

Fig. 11.3 Transition diagrams pre-measure (*left*) and post-measure (*right*) of Teacher initiation (T), Student task-related utterance (S), Teacher's stimulating response (T st), and Teacher's non-stimulating response (T N-st)

reflection). A significantly different interaction pattern is found between pre- and post-measure ($p < 0.01$) in that during the post-measure an initiation question of the teacher is often (in 77 % of the cases) directly followed by a task-related student utterance. Next, a student utterance is most often followed by a stimulating follow-up of the teacher. This seems to indicate better attuned interactions, i.e., stimulating interactions, possibly emerging into higher levels of student complexity.

Conclusion and Discussion

From a *content-based perspective*, the surplus value of a complex dynamic systems approach was illustrated by analyzing the (effect of) the Video Feedback Coaching program for teachers intervention, in which complexity properties were intertwined in design, data collection, and analysis.

When looking at the aggregated and static data, the results showed a positive intervention effect on the macro level of students' science performance. The question arose about the practical significance of this result. An average increase of 1 complexity level seemed trivial. The effect size ($d = 1.6$) showed that this effect can be considered very large. However, this number does not provide practical tools for teachers. By using a process-based intervention study the surplus value of applying the properties became clear:

1. By incorporating time serial aspects of change, the intervention effect could be further explained. The average trajectory of all classrooms over several lessons (dynamic) showed a rather stable level during the intervention. The effect of the intervention on students' performance only became apparent at post-measure.
2. By focusing on intra-individual variability, however, it became clear that the average trajectory underestimated the variability present in individual trajectories. Half of the classrooms showed a rather stable trajectory, while the other half represented great variability. None of the groups showed a clear positive intervention effect on students' scientific reasoning level *during* the intervention sessions. However, previous research indicated that before a new state (i.e., higher level of performance) can be reached, a period of "increased variability" appears (Bassano & Van Geert, 2007; Van Der Steen et al., 2014; Van Geert & Van Dijk, 2002). These suggestions can be further analyzed by focusing on micro-dynamic processes in all lessons, in order to find out whether there is more variability leading to a new state at the micro level during the lessons of the intervention period. Another explanation for the, in this case rather high, variability might be found by focusing on the lesson characteristics. When a teacher provides a lesson mainly focussing on following the steps on a worksheet, a different interactional quality might be expected compared to a lesson in which students have more degrees of freedom to experiment. Note that the transactional nature might be used to further interpret this qualitative finding.

3. By examining the transactional nature, it became apparent that the higher performance seems to be achieved by a mutual investment of teacher and students and that a change in interaction patterns seems to underlie this phenomenon. The representative case showed that an increase in students' understanding is accompanied by a change in interactional quality and that the students' scientific reasoning level fluctuates in interaction with the teacher. During the post-measurement, teacher and students seem more attuned to each other, in that a teacher's question is twice as often followed by a student answer compared to the pre-measurement. Students seem more capable of using complex terms to express their thinking processes, as is expressed in the higher complexity scores. In addition, during post-measurement, the student utterance is only followed by a stimulating response, while during pre-measure, non-stimulating utterances were apparent. Based on the micro-dynamic data, we therefore suggest that the higher performance during the post-measurement can be explained by interactions of higher quality in which the teacher poses more stimulating questions and that the students reason on higher scientific reasoning levels. The point of this type of analysis is not to pretend that these percentages apply to the population, as an average level. We aimed to depict a technique of representation that shows the time serial nature of the process. It goes without saying that the structure of these processes may be quite different for one case in comparison to another, but the nature of the representation, in terms of a transition diagram, in principle applies to all possible forms of interaction in classrooms. By choosing a different way of representing the interaction in the classroom, namely by means of these transition diagrams, the emphasis which is traditionally put on static measures, is now replaced by a dynamic representation, which in some cases may be of quite considerable complexity. Especially for teachers, the latter might be a more accurate reflection of the teacher's real time experiences as teachers are "aware" —usually without being familiar with the technical terms— that they are working within a complex dynamic system.

To summarize, the surplus value of the analysis is that it illustrates how a complex dynamic systems approach can be used to describe the processes underlying static group-based educational intervention effects, and provide information about the quality of that intervention. By using a process-based methodology, we were able to show that average results can be deepened by focusing on several complexity properties. We suggested answers to the question of why the VFCt intervention worked and why it seemed to work better during some lessons compared to other lessons within one classroom (i.e., type of questions, attuned interactions, using active participation during experiments versus classical experiment lessons). In addition, insight was provided into the actual changes during lessons and how interaction proceeded. This information cannot be found in conventional longitudinal studies, but are essential for teachers as this might more accurately reflect what they experience during their lessons and gives insight into how teachers can optimize their lessons—compared to standard evaluations.

Of course, when assessing the effectiveness of an intervention the use of a control group will primarily provide information about differences between the actual processes; especially the micro-process differences (see for instance Wetzels et al., 2015). Veerman and van Yperen (2007) state that the use of a control group is a prerequisite for analysis of the effectiveness. Therefore, the next step is to analyze classrooms that did not participate in the VFCt, but did provide science and technology lessons (Van Vondel et al., 2015).

From a *methodological point of view*, we would like to make a distinction between "hard" complex dynamic systems research and "soft" complex dynamic systems research in education. The distinction might be somewhat exaggerated and is rather a matter of degree, but we think it is important to discuss it in order to put much of the complex dynamic systems research that is currently being done in education in the right perspective.

By "hard" complex dynamic systems research, we mean the research that focuses on typical complex dynamic systems properties and which is based on very dense time series. Examples are studies of attractors and discontinuities, for instance by means of cusp catastrophe models (Van Der Maas & Molenaar, 1992), or studies of the statistical structure of time series revealing properties such as pink noise in rower's coordination of ergometer strokes (Den Hartigh, Cox, Gernigon, Van Yperen, & Van Geert, 2015) or studies using techniques such as recurrence quantification analysis that try to reconstruct the complexity of the state space that underlies the attractors of the system (Wijnants, Bosman, Hasselman, Cox, & Van Orden, 2009).

By "soft" complex dynamic systems research, in contrast, we mean educational research inspired by basic, qualitative features of a complex dynamic systems view on education and which is rooted in educational practice, as the VFCt. Some examples which would typically qualify as "soft" complex dynamic systems research are presented by Steenbeek, et al. (2012): research on learning that focuses on individual trajectories and on intraindividual variability, on the transactional and iterative nature of the teaching-learning process and on the relationship between the short-term time scale of learning activities and the long-term time scale of development. It is a kind of research that describes how such patterns are self-sustaining and hard to change, i.e., tends to show considerable resistance to change and thus have the qualitative properties of attractor states.

Scientific implications for intervention studies: Especially evaluation studies of —applied— educational interventions are fruitful areas for a "soft" complex dynamic systems approach. As performance is usually constructed in interaction between a more knowledgeable partner and a student (Steenbeek & Van Geert, 2013; Van De Pol et al., 2011), observational classroom studies provide rich information. Analyses on the micro-level show whether the effect of an intervention can be found on the level where interventions focus at, in this case on interactions of higher quality. For a complete understanding of the process of teaching students a particular way of reasoning, an intensive study of a teacher's —in combination with the students'— behavior over several lesson will reveal important insights. Focusing on "how" an

intervention works is a way of describing why one state changes into another, and in fact implies a way of describing what can be done to make the state change into another one (Van Geert & Steenbeek, 2005). Furthermore, the case study findings can be supported by findings of a multiple case study. These findings can then be used to generalize findings and by that strengthen evidence-based practice.

Practical implications: The results of process analysis can be used in two different ways, as both scientific and practical purposes can be highlighted. First, the results add to fundamental knowledge about how scientific reasoning skills are (co-) constructed in real-time (Meindertsma et al., 2014) and how the effect of a teaching intervention emerges during actual science and technology lessons. Second, the results can be used for educational purposes. This approach provides accessible practice-based tools for best practice, or perhaps more importantly, familiar examples which can be used for (in-service) teacher professionalization (Wetzels et al., 2015). The micro-dynamic analysis might map the most interesting information for educational practitioners as it yields practice-based results.

Further analyses: An important next step for the study of interventions is to map the teacher–student interactions of individual teachers in order to study whether interindividual variability can be further explained on the micro-level (Van Vondel, Steenbeek, Van Dijk, & Van Geert, in preparation). The analyses of the empirical example as presented in this paper may be not more than only the first steps towards a complex dynamic systems approach. More information can be extracted by repeating similar analyses for teacher variables, by focusing on all lessons of individual teachers, by comparing micro and macro findings, or by comparing two extreme cases on the micro level (e.g., Steenbeek et al., 2012).

To conclude, interventions should be studied as emerging processes on various, intertwined time scales taking place in individual cases, and not as isolated causal factors, with an intrinsic effectiveness, applying to a specific population category. We, therefore, stress the importance of using variables that capture the transactional character of interventions, specifically when they are aimed at improving interaction patterns in the naturalistic classroom situation. For future research we like to state that it is essential to look more closely at what the intervention is aiming at and what the role of the immediate context/proximal factors are in this process. When more understanding is gained about what happens during the intervention, for instance about stability or change in interaction patterns, intervention programs can be specifically attuned to supporting high quality interaction patterns in the classroom and students can thus be stimulated to perform optimally.

References

Bassano, D., & Van Geert, P. L. C. (2007). Modeling continuity and discontinuity in utterance length: A quantitative approach to changes, transitions and intra-individual variability in early grammatical development. *Developmental Science, 10*(5), 588–612.

Bentley, M. L. (1995). Making the most of the teachable moment: Carpe diem. *Science Activities, 32*(3), 23–27.

Boelhouwer, M. D. (2013). *Tussen weerstand en weerbaarheid, en andere recepten; Een effectevaluatie van het WIBO-lesprogramma met behulp van vragenlijsten, dagboeken en observaties* (Between resistance and empowerment; the evaluation of the WIBO-program prevention program). Groningen, The Netherlands: University of Groningen.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness, 1*, 237–264. doi: 10.1080/19345740802328273.

Bos, J., & Steenbeek, H. W. (2009). *Mediacoder; software voor het coderen van video- en audiomaterialen*. Groningen: Internal publication, IDP Department, University of Groningen.

Chin, C. (2006). Classroom interaction in science: Teacher questioning and feedback to students' responses. *International Journal of Science Education, 28*(11), 1315–1346. doi: 10.1080/09500690600621100

Cumming, G. (2014). The new statistics why and how. *Psychological Science, 25*(1), 7–29. doi: 10.1177/0956797613504966

De Groot, A. D. (1994). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen*. Assen: Van Gorcum.

Den Hartigh, R. J. R., Cox, R. F. A., Gernigon, C., Van Yperen, N. W., & Van Geert, P. L. C. (2015). Pink noise in rowing ergometer performance and the role of skill level. *Motor Control, 19*, 355–369. doi: 10.1123/mc.2014-0071

Den Hartigh, R. J. R., Gernigon, C., Van Yperen, N. W., Marin, L., & Van Geert, P. L. C. (2014). How psychological and behavioral team states change during positive and negative momentum. *PLoS One, 9*(5), e97887. doi: 10.1371/journal.pone.0097887

Ensing, A., Van der Aalsvoet, D., Van Geert, P.L.C., & Voet, S. (2014). Learning potential is related to the dynamics of scaffolding. An empirical illustration of the scaffolding dynamics of five year olds and their teacher. *Journal of Cognitive Education and Psychology, 13*(3), 1–18. doi: 10.1891/1945-8959.13.3.375

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*, 477–531. doi: 10.1037/0033-295X.87.6.477.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought, and emotion. In R. M. Lerner (Ed.) & W. Damon (Series ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 313–399). New York: Wiley.

Fogel, A. (2011). Theoretical and applied dynamic systems research in developmental science. *Child Development Perspectives, 5*(4), 267–272. doi: 10.1111/j.1750-8606.2011.00174.x.

Fredrickson, B. L. (2015). The broaden-and-build theory of positive emotions. *Philosophical Transactions, 359*(1449), 1367–1377. doi: 10.1098/rstb.2004.1512

Fukkink, R. G., Trienekens, N., & Kramer, L. J. (2011). Video feedback in education and training: Putting learning in the picture. *Educational Psychology Review, 23*(1), 45–63.

Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education, 86*(5), 693–705. doi: 10.1002/sce.10039

Granott, N., & Parziale, J. (2002). Microdevelopment: A process-oriented perspective for studying development and learning. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 1–28). New York, NY: Cambridge University Press.

Hock, M. F., Schumaker, J. B., & Deschler, D. D. (1995). Training strategic tutors to enhance learner independence. *Journal of Developmental Education, 19*, 18–26.

Hollenstein, T., & Lewis, M. D. (2006). A state space analysis of emotion and flexibility in parent-child interactions. *Emotion, 6*(4), 656. doi: 10.1037/1528-3542.6.4.656

Hood, G. (2004). Poptools [Computer software]. Pest Animal Control Co-operative Research Center.

Hyun, E., & Marshall, J. D. (2003). Teachable-moment-oriented curriculum practice in early childhood education. *Journal of Curriculum Studies, 35*(1), 111–127. doi: 10.1080/00220270210125583

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112–133. doi: 10.1177/1558689806298224

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—Significance tests are not. *Theory and Psychology, 22*(1), 67–90. doi: 10.1177/0959354311429854

Lehmann, A. C., & Gruber, H. (2006). Music. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 457–470). New York, NY: Cambridge University Press.

Lewis, M. D. (1995). Cognition-emotion feedback and the self-organization of developmental paths. *Human Development, 38*(2), 71–102. doi: 10.1159/000278302

Lewis, M. D. (2002). Interacting time scales in personality (and cognitive) development: intentions, emotions, and emergent forms. In N. Grannot & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 183–212). New York, NY: Cambridge University Press.

Lichtwarck-Aschoff, A., van Geert, P. L. C., Bosma, H., & Kunnen, S. (2008). Time and identity: A framework for research and theory formation. *Developmental Review, 28*(3), 370–400. doi: 10.1016/j.dr.2008.04.001.

Meindertsma, H. B., Van Dijk, M. W. G., Steenbeek, H. W., & Van Geert, P. L. C. (2012). Application of skill theory to compare scientific reasoning of young student in different tasks. *Netherlands Journal of Psychology, 67*, 9–19.

Meindertsma, H. B., van Dijk, M. W. G., Steenbeek, H. W., & van Geert, P. L. C. (2014). Assessment of preschooler's scientific reasoning in adult–child interactions: What is the optimal context? *Research in Science Education, 44*, 215–237. doi: 10.1007/s11165-013-9380-z.

Mortenson, B., & Witt, J. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review, 27*(4), 613–627.

Noell, G. H., Witt, J., Slider, N., Connell, J., Gatti, S., Williams, K., et al. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review, 34*(1), 87–106.

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching, 47*(4), 422–453.

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*, 177–196. doi: 10.1007/s10972-014-9384-1.

Parziale, J., & Fischer, K. (1998). The practical use of skill theory in classrooms. In R. Sternberg & W. Williams (Eds.), *Intelligence, instruction, and assessment: Theory into practice* (pp. 95–110). Mahwah: Lawrence Erlbaum Associates Publishers.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal, 48*(4), 996–1025. doi: 10.3102/0002831211410864

Pintrich, P. R. (2000). The role of goal orientation in self regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of selfregulation* (pp. 451–502). San Diego: Academic.

Reinke, W. M., Sprick, R., & Knight, J. (2009). Coaching classroom management. In J. Knight (Ed.), *Coaching: Approaches & perspectives* (pp. 91–112). Thousand Oaks: Corwin Press.

Rosmalen, J. G., Wenting, A. M., Roest, A. M., de Jonge, P., & Bos, E. H. (2012). Revealing causal heterogeneity using time series analysis of ambulatory assessments: Application to the

association between depression and physical activity after myocardial infarction. *Psychosomatic Medicine, 74*(4), 377–386. doi: 10.1097/PSY.0b013e3182545d47

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education, 27*(2), 259–267. doi: 10.1016/j.tate.2010.08.009

Şimşek, P., & Kabapinar, F. (2010). The effects of inquiry-based learning on elementary students' conceptual understanding of matter, scientific process skills and science attitudes. *Procedia - Social and Behavioral Sciences, 2*(2), 1190–1194. doi: 10.1016/j.sbspro.2010.03.170.

Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*(8), 343–348. doi: 10.1016/S1364-6613(03)00156-6

Steenbeek, H. W., Jansen, L., & Van Geert, P. L. C. (2012). Scaffolding dynamics and the emergence of problematic learning trajectories. *Learning and Individual Differences, 22*(1), 64–75. doi: 10.1016/j.lindif.2011.11.014

Steenbeek, H. W., & Van Geert, P. L. C. (2005). A dynamic systems model of dyadic interaction during play of two student. *European Journal of Developmental Psychology, 2*(2), 105–145. doi: 10.1080/17405620544000020

Steenbeek, H. W., & Van Geert, P. L. C. (2013). The emergence of learning-teaching trajectories in education: A complex dynamic systems approach. *Nonlinear Dynamics, Psychology, and Life Sciences, 17*(2), 233–267.

Steenbeek, H. W., & Van Geert, P. L. C. (2015). Assessing young student's learning and behavior in the classroom; a complexity approach. In G. M. Van der Aalsvoort (Eds.), *International handbook on early childhood education, Volume 2—Curriculum and assessment*.

Steenbeek, H. W., Van Geert, P. L. C., & Van Dijk, M. W. G. (2011). The dynamics of student's science and technology talents: A conceptual framework for early science education. *Netherlands Journal of Psychology, 66*(3), 96–109.

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. doi: 10.4300/JGME-D-12-00156.1

Thelen, E. (1992). Development as a dynamic system. *Current Directions in Psychological Science, 1*(6), 189–193. doi: 10.1111/1467-8721.ep10770402.

Van de Pol, J., Volman, M., & Beishuizen, J. (2011). Patterns of contingent teaching in teacher–student interaction. *Learning and Instruction, 21*(1), 46–57. doi: 10.1016/j.learninstruc.2009.10.004

Van der Maas, H., & Molenaar, P. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99*, 395–417. doi: 10.1037/0033-295X.99.3.395

Van der Steen, S., Steenbeek, H. W., Van Dijk, M. W. G., & Van Geert, P. L. C. (2015). How to characterize the development of student's understanding of scientific concepts: A longitudinal microgenetic study. Manuscript submitted for publication.

Van der Steen, S., Steenbeek, H. W., Van Dijk, M. W. G., & Van Geert, P. L. C. (2014). A complexity approach to student's understanding of scientific concepts: A longitudinal case study. *Learning and Individual Differences, 30*, 8–91. doi: 10.1016/j.lindif.2013.12.004

Van der Steen, S., Steenbeek, H. W., Wielinski, J., & Van Geert, P. L. C. (2012). A comparison between young students with and without special needs on their understanding of scientific concepts. *Educational Research International, 2012*, 1–12. doi: 10.1155/2012/260403.

Van der Steen, S., Steenbeek, H. W., & Van Geert, P. L. C. (2012). Using the dynamics of a person-context system to describe student's understanding of air pressure. In H. Kloos, B. J. Morris, & J. L. Amaral (Eds.), *Current topics in student's learning and cognition* (pp. 21–44). Rijeka: InTech.

Van Geert, P. L. C. (1994). *Dynamic systems of development: Change between complexity and chaos*. New York, NY: Harvester Wheatsheaf.

Van Geert, P. L. C. (1997). Time and theory in social psychology. *Psychological Inquiry, 8*, 143–153.

Van Geert, P. L. C. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review, 105*(4), 634. doi: 10.1037/0033-295X.105.4.634-677

Van Geert, P. L. C. (2003). Dynamic systems approaches and modeling of developmental processes. In J. Valsiner & K. J. Conolly (Eds.), *Handbook of developmental psychology* (pp. 640–672). London: Sage.

Van Geert, P. L. C. (2004). Dynamic modelling of cognitive development: Time, situatedness and variability. In A. Demetriou & A. Raftopoulos (Eds.), *Cognitive developmental change: Theories, models and measurement* (pp. 354–379). Cambridge, MA: Cambridge University Press.

Van Geert, P. L. C., & Fischer, K. W. (2009). Dynamic systems and the quest for individual-based models of change and development. In J. P. Spencer, M. S. C. Thomas, & J. McClelland (Eds.), *Toward a unified theory of development? Connectionism and dynamic systems theory reconsidered* (pp. 313–336). Oxford: Oxford University Press.

Van Geert, P. L. C., & Steenbeek, H. W. (2005). Explaining after by before: Basic aspects of a dynamic systems approach to the study of development. *Developmental Review, 25*(3–4), 408–442. doi: 10.1016/j.dr.2005.10.003

Van Geert, P. L. C., & van Dijk, M. W. G. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development, 25*(4), 340–374.

Van Vondel, S., Steenbeek, H. W., Van Dijk, M. W. G., & Van Geert, P. L. C. (2015). The effects of Video Feedback Coaching for teachers on elementary students' scientific knowledge. Manuscript submitted for publication.

Van Vondel, S., Steenbeek, H.W., Van Dijk, M.W.G., & Van Geert, P.L.C. (in preparation). Ask Don't Tell; Improving science education by focusing on the co-construction of scientific understanding.

Veerman, J. W., & van Yperen, T. A. (2007). Degrees of freedom and degrees of certainty: A developmental model for the establishment of evidence-based youth care. *Evaluation and Program Planning, 30*(2), 212–221.

Velicer, W. (2010). *Applying idiographic research methods: Two examples*. In 8th International Conference on Teaching Statistics, Ljubljana, Slovenia.

Vygotsky, L. (1986). *Thought and language*. Cambridge: Cambridge University Press.

Wetzels, A. F. M., Steenbeek, H. W., & Van Geert, P. L. C. (2015). Primary science teaching: Behavior of teachers and their pupils during and after a coaching program. Manuscript submitted for publication.

Wetzels, A. F. M., Steenbeek, H. W., & Van Geert, P. L. C. (in press). A complexity approach to the effectiveness of an intervention for lower grade teachers on teaching science. *An International Journal of Complexity and Education*.

Wijnants, M. L., Bosman, A. M., Hasselman, F., Cox, R. F., & Van Orden, G. C. (2009). 1/f scaling in movement time changes with practice in precision aiming. *Nonlinear Dynamics, Psychology, and Life Sciences, 13*(1), 79.