SECONDARY STUDENTS' CONSIDERATIONS OF VARIABILITY IN MEASUREMENT ACTIVITIES BASED ON AUTHENTIC PRACTICES

ADRI DIERDORP
Utrecht University
adridierdorp@gmail.com

ARTHUR BAKKER
Utrecht University
a.bakker4@uu.nl

DANI BEN-ZVI
The University of Haifa
dbenzvi@univ.haifa.ac.il

KATIE MAKAR
The University of Queensland
k.makar@uq.edu.au

ABSTRACT

Measurement activities were designed in this study on the basis of authentic professional practices in which linear regression is used, to study considerations of variability by students in Grade 12 (aged 17–18). The question addressed in this article is: In what ways do secondary students consider variability within these measurement activities? Analysis of students' reasoning during these activities in one classroom (N = 13) suggests that students considered variability in four ways: noticing and acknowledging variability, measuring variability, explaining variability, and using investigative strategies to handle variability. We conclude that the measurement tasks based on authentic professional practices helped students to reason with relevant aspects of variability. Finally, we discuss curricular and research implications.

Keywords: *Statistics education research; Authentic professional practice; Linear regression; Statistical reasoning.*

1. INTRODUCTION

Variability is omnipresent. Variability is the phenomenon that something is apt or liable to change (Reading & Shaughnessy, 2004). Wild and Pfannkuch (1999) stress that "variability affects all aspects of life and everything we observe. No two manufactured items are identical, no two organisms are identical or react in identical ways." (p. 235). In the human quest for certainty and knowledge (Dewey, 1929), variability makes description, analysis, and conclusion a challenge. For example, a sports physiologist who measures a person's heart rate and uses a formula to describe this person's physical fitness faces various sources of variability and can respond to this variability in different ways. Variability in this scenario can be due to measurement error but also to natural variability

in a person's heart rate. If a heart rate measurement is very high, she should notice the unusualness of it, try to explain the high value, be aware of possible sources of variability, and find ways to assess the measurement accuracy.

Filtering such messages from noisy data is a key aspect of statistical thinking (Wild & Pfannkuch, 1999). A signal in data does not just refer to central tendency, but can also describe stability in variability measured with a range, interquartile range, or standard deviation; a signal can also be the shape of a distribution (Bakker, 2004) or a trend (e.g., Fitzallen, 2012). Variability is a thus a broader concept than spread or variation (Shaughnessy, 2007). Statistics is more than the science of variability (MacGillivray, 2004), but also the science of identifying and investigating stability or signals in the noise (Konold & Pollatsek, 2002).

Despite the importance of variability and students' difficulty understanding it (Ben-Zvi, 2004; Garfield & Ben-Zvi, 2005), many statistics curricula have a narrow focus on identifying and measuring centres of data sets (Sorto, 2006) rather than considering variability. As Reading and Shaughnessy (2004) pointed out "Students' current lack of understanding of the nature of variability in datfxa may be partly due to the lack of emphasis of variability in our traditional school mathematics curriculum and textbooks" (p. 203). For example, in a review of Dutch secondary school mathematics textbooks with statistics chapters, we found that none of them made explicit reference to variability or variation even though measures of spread such as IQR and standard deviation are introduced. Hjalmarson, Moore, and delMas (2011) stress that the lack of tasks that require students to measure variability may impede their understanding of variability. Even when Hjalmarson et al. (2011) found a few examples of tasks in engineering textbooks that provoked students to measure variability, these tasks were disconnected from a real-world context.

While seeking real world investigational contexts that may support reasoning in the presence of variability, we have explored the idea of engaging students in measurement tasks (Enderson, 2003) based on authentic professional practices (Prins, 2010; Westbroek, Klaassen, Bulte, & Pilot, 2010). As explained in the next section, we assumed that in such activities students would reason with variability in rich ways. The goal of this article is to contribute to knowledge of *how students can learn to consider variability in measurement tasks based on authentic professional practices.*

2. THEORETICAL BACKGROUND

2.1. VARIABILITY

As variability is a multifaceted concept, multiple aspects should be promoted and evaluated in teaching and learning statistics (Reid & Reading, 2008). In their analysis of statisticians' thinking, Wild and Pfannkuch (1999) distinguished four ways in which statistical experts consider variability:

1. Noticing and acknowledging variability
2. Modeling or measuring variability for the purpose of predicting, explaining and controlling
3. Explaining and dealing with variability
4. Using investigative strategies to handle variability

Variability was extensively discussed by Wild and Pfannkuch (1999) in their study of how statisticians think (Figure 1). They stressed the importance of considering variability

when exploring data. Besides the unexplained real variability that is typical for a system, they distinguish *induced variability* (p. 235) caused by the various ways of collecting data. They mention measurers and devices affecting measurements as a reason for induced variability, but also sampling and accidents caused by the data collection or the method itself.

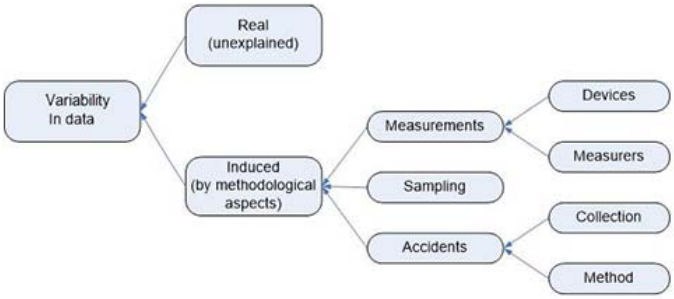

Figure 1. Sources of variability in data, based on Wild and Pfannkuch (1999)

To understand variability, one needs to reason with statistical and contextual worlds in relation to each other (Ben-Zvi & Aridor, 2016). However, education typically focuses on only one of these components. Statistics education (as part of mathematics education) can often focus on techniques that neglect authentic application or cross-curricular content. Conversely, content areas such as science education may focus on scientific content, but pay little attention to statistical techniques (Estepa & Sanchez-Cobo, 1998; Reading & Shaughnessy, 2004). The outcome of this discrepancy is that students are inexperienced in applying statistical concepts to contextually rich content or authentic tasks in science-related contexts. Makar and Confrey (2007) argue that students who engage in statistical inquiry with a compelling purpose, such as experiences with authentic data, gain a deeper understanding of data analysis and the context itself. One challenge is to design learning activities that connect statistical techniques and ideas in authentic contexts that are rich in scientific content. This can help students to notice, acknowledge, and deal with variability, and seek meaningful explanations for variability in applications of science. We focused this study on variability in the form of deviations (residuals) from a regression line, which are partly due to real (natural) deviations and partly due to measurement error.

2.2. MEASUREMENT LEADING TO THE INVESTIGATION OF REAL-WORLD DATA

Measurement is the assignment of numbers with units to objects or events (Pedhazur & Pedhazur Schmelkin, 1991) and can be described as ordering our surrounding world through numbers to better understand that world (Adams & Harrell, 2003; Buys & de Moor, 2005). Measurement has been gaining greater importance in society: To participate successfully in modern society, it is important that students learn both to measure various phenomena in their environment and how to analyse the resulting measurements (Franklin et al., 2005; Gooya, Khosroshahi, & Teppo, 2011; Lehrer & Kim, 2009).

An advantage of involving students in measurement activities is that it invites them to make connections between the real world and the world of data, and thus learn to see that measurement cannot be absolutely accurate (Rabinovich, 2005). In other words, generating real-world data can help students to see measurements as estimates with natural variability and measurement errors. Activities with a discussion of the measurement process and the

resulting data can increase students' understanding of the nature and importance of measurement (Moore, 1990).

To understand measurement data, students need considerable contextual background, including knowledge of the phenomenon measured and the measurement procedure. Graphical representation of the data can help students to see the variability around the regression line and develop their understanding of variability (delMas & Liu, 2005). Such representations allow students to see shapes or trends in data, and make predictions. Research on graphing (e.g., Roth & Bowen, 2003) suggests that students should be involved in measurement processes in order to interpret resulting data. Even professional scientists require much contextual background to interpret graphs, and if not familiar with the data generation process, they may find it difficult to read graphs in their own discipline. In fact, Roth and Bowen (2003) recommend that experience with research and participation in graphing practices was more important for correct graph interpretation than exposure to increasingly complex graphs. We therefore chose to involve students in measurement activities that stress the importance of contextual background when graphing real data. Inspired by research in science education (Prins, 2010; Westbroek et al., 2010), we chose to base the design of activities on authentic professional practices.

In this article we define an authentic professional practice (AuPP) as a patterned purposeful activity of professionals working on a problem that is exemplary for their profession. In science education, learning activities based on AuPPs can offer students meaningful contextual links to abstract concepts (Lee & Butler-Songer, 2003). The activities based on AuPPs have to be to simplified or modified to make them useful in an educational setting. For example, Dierdorp, Bakker, Eijkelhof, and van Maanen (2011) based their design of statistics activities on an AuPP of monitoring the height of dykes in the Netherlands, in which students used their contextual knowledge to make sense of variability in real data. What became clear in this earlier study is that variability needed even more attention in design and teaching, hence the study reported here.

2.3. RESEARCH QUESTION

The current article reports on research in which students in grade 12 (17–18 years old) reasoned with variability when they engaged in a simplified AuPP of a sports physiologist. To analyse aerobic and anaerobic respiration, they measured heart rates under increasing physical effort and applied regression techniques concerning variability in data in order to determine the ideal heart rate (threshold point) at which the working of muscles turns from aerobic to anaerobic metabolism. The idea underlying our research was that measurement activities based on suitable AuPPs could support students in reasoning with variability in different ways. To evaluate this idea, we asked the following research question: *In what ways do secondary students consider variability within measurement activities based on authentic professional practices?*

3. METHOD

3.1. RESEARCH SETTING

The data presented here stem from the first author's PhD research project (Dierdorp, 2013) investigating how students can learn the key statistical concept of regression in multidisciplinary contexts, experiencing the links between mathematics and the natural sciences. The overall study was based on design-based research, which involved an iterative design process (Barab & Squire, 2004; Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006) consisting of six research cycles. Each cycle included the design of a

hypothetical learning trajectory (Simon, 1995), a teaching experiment of about twenty lessons to implement and assess the instructional unit and students' learning, analysis of classroom data, and revision of the learning trajectory. This paper reports on the analysis of the fifth research cycle, which focused on students' reasoning with variability.

As we were unable to find existing measurement activities in secondary school statistics based on authentic measurements by professionals, we designed an instructional unit ourselves. There was little educational research to draw on as most of the research on measurement has been carried out in primary education, focusing on spatial measurement (e.g., Lehrer, 2003). Most research concerns relatively straightforward measurement of parameters such as length and volume with simple technologies such as rulers and measuring jugs (e.g., Smith III, Van den Heuvel-Panhuizen, & Teppo, 2011). What comes closest to what we envisioned is the work by Lehrer, Kim, and Schauble (2007) in primary science education, which incorporates modeling and data analysis (see also English, 2009), topics related closely to variability and measurement.

In the design of the learning trajectory, we searched for suitable AuPPs that contained measurement activities in which professionals use regression lines that could be adapted for students in grades 11 and 12. We also wanted students to appreciate the AuPPs and identify with the professionals in ways that coherently embrace mathematics and the natural sciences. It was particularly important for us that there would be at least one AuPP-based activity in which students could perform a measurement experiment. These considerations led us to the practices of sports physiologists identifying the best training procedure for their clients. Based on typical practices, we designed two measurement activities described in more detail in Section 3.3.

3.2. PARTICIPANTS

Thirteen students, seven boys and six girls, from an affluent school in the Netherlands took part in this study. They were in the beginning of Grade 12 of the pre-university track, which is attended by the top 15% of academically achieving Dutch students. The first author (T) taught these students at his own school in a small city. The designed instructional unit was entitled "Statistics as a Bridge between Mathematics and the Natural Sciences" and was part of their school subject "Nature, Life, and Technology" (Eijkelhof & Krüger, 2009). The students participated in classroom discussions and worked in pairs and small groups. They were asked to reason and explain their actions more than they were used to in other school subjects.

3.3. THE MEASUREMENT LEARNING ACTIVITIES

In this section we describe the two measurement activities, which spanned three lessons each (50 minutes per lesson). These measurement activities aimed to involve students in reasoning with variability in informal ways in relation to regression to prepare for the learning of formal regression techniques in subsequent lessons. In the first measurement activity the students were asked to perform heart rate measurements and use a given formula to quantify physical fitness. In the second measurement activity they were asked to construct a regresson line using their own measurements of heart rates under increasing physical effort.

Measurement Activity 1 (MA1): Measuring Fitness MA1, which consists of six tasks with several subtasks, concerns the measurement of physical fitness. Professional sports physiologists regularly use measurements and regression techniques in their advice about the best training for their clients; in particular, accurate measurements and suitable

statistical techniques are needed when they want to determine the physical fitness of a person and assess their potential and risks. We assumed that students could engage easily with this context because many of them care about their physical fitness and do some sport themselves. They presumably have some prior knowledge of this AuPP and see the purpose of it. This would help them see the usefulness of what they learned (cf., Lijnse & Klaassen, 2004).

The design aim of MA1, on the measurement of physical fitness, was to stimulate students to reason with variability. We expected to achieve this goal by allowing students to perform their own measurements and compare these with an existing formula. We assumed that suitable AuPPs constitute rich contexts that are meaningful for students, which would make it easier for them to reason with variability (Cobb & Moore, 1997) and be motivated to learn (Dierdorp, Bakker, van Maanen, & Eijkelhof, 2014). We conjectured that a) the authentic data would show enough "noise" to urge students to notice and acknowledge variability when they interpreted the data (Konold & Pollatsek, 2002); b) they would understand that the relation between physical effort and heart rate could be predicted with a regression line by estimating parameters or correlation; c) they would explain the noise by sources of variability; and d) they would use investigative strategies such as representing the data with graphs to seek ways to interpret the variability.

Heart rate increases with increased physical effort, but this happens less rapidly with people who are in good physical fitness than with people who are less physically fit. In addition, people who train regularly recover more rapidly after physical effort (heart rate becomes normal again). Researchers have designed suitable tests to quantify physical fitness by measuring heart rates. The Ruffier-Dickson test (Paulet, Gratas, Dassonville, & Rochcongar, 1981) uses heart rate frequencies at three relevant moments in a physical exercise to determine physical fitness. In MA1, partly presented in Figure 2, students were asked to use and discuss this test. We expected them to demonstrate the four ways that statistical experts consider variability (see Section 2.1). After the completion of the heart rate measurements, students were asked to calculate their Ruffier-Dickson Index (RDI) and to reason with variability.

The students were allowed to choose the method of measuring heart rate. Some used their own hands and counted the heart rate, others used a sphygmometer with a heart rate monitor incorporated. In the left picture of Figure 3 a knee bending male student is shown while another female student is making notes. In the right picture of Figure 3 the female student is placing the sphygmomanometer.

In order to explain the effect of H1, H2, and H3, on the RDI, the students were also asked to find the relation between H1 and H2. We expected that students would use investigative strategies for finding a trend in their data by using regression lines.

Task 2a

Introduction

In this task you will measure the heart rate frequency (HRF) the same way as professional sports physiologists do. The task concerns three measurements. First a measurement at rest. Secondly, a measurement after knee bends. Third, a measurement at rest again. Work in small groups. Every student will take the test (testee) and perform the measuring at least once (measurer). The following instructions describe how to take the measurements.

Heart Rate Measurement

The testee needs to sit quietly for about one minute before starting the measurement process. Measure the heart rate (number of beats) of the testee at rest for 15 seconds and convert it to beats per minute. We call this resting heart rate **H1**.
Always measure the heart rate with your middle finger (possibly joined by the index finger). The artery is on the side of your thumb.
Have someone else perform the measurement, so the testee doesn't need to keep an eye on the time.

Next, the testee does 30 deep knee bends in about 45 seconds. The back remains straight and the feet must keep contact with the ground. Each time your fingertips should touch the ground.
Measure the heart rate (number of beats) directly afterwards, for 15 seconds and convert it to beats per minute (**H2**). Sit quietly again. One minute later measure your heart rate again for 15 seconds and convert it to one minute (**H3**).
An indication of your physical fitness can be calculated from these measured values using the Ruffier-Dickson Index (RDI) formula. This index is frequently abbreviated as RDI and is defined as:

$$RDI = \frac{H2 - 70 + 2 \cdot (H3 - H1)}{10}$$

Translation from RDI to a qualitative indication of physical fitness:

RDI	**Physical Fitness**
Below or equal to 0	excellent
between 0 and 3	very good
between 3 and 6	good
between 6 and 8	moderate
above 8	bad

Figure 2. Task 2a from Measurement Activity MA1 (translated from Dutch)

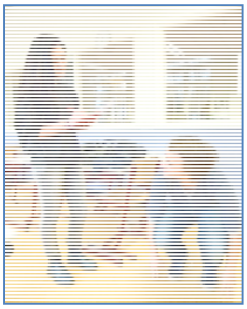

Figure 3. Students collect data

Measurement Activity 2 (MA2): Identifying a Suitable Sports Program MA2 consists of four tasks with several subtasks. One task concentrated on presenting and analysing data collected by students performing the Conconi Test (Conconi, Ferrari, Ziglio, Droghetti, & Codeca, 1987), which measures the threshold heart rate frequency (HRF) at which the

muscles switch from aerobic to anaerobic combustion. Despite recent studies that have shown the Conconi Test has limited accuracy levels, we decided that it is a good option to offer students because it is still used and is suitable for students to reason with variability in relation to linear regression.

The design aim for MA2 was for students to reason with various types of variability. As argued in Section 2.2, it would be important for them to collect and investigate data themselves in order to enrich their ability to interpret the resulting graphs (Roth & Bowen, 2003). These important opportunities for students and the mentioned design aim made us base MA2 on a practice of sports physiologists who identify the best training program for clients. Just like athletes, students had to measure their HRF with a heart rate monitor while increasing the intensity of their effort (selected power at the treadmill). It is important for athletes to stay within the aerobic area otherwise the muscles produce lactic acid. Training in the aerobic area under the threshold HRF from aerobic to anaerobic will prevent a decrease of this threshold and prevent muscle problems. According to Gellish et al. (2007), when people increase their efforts during their training session, HRF increases proportionally with the physical effort. If the effort exceeds a certain point, the linear proportionality will disappear and the HRF will approach the peak heart rate (Figure 4). The upper bound where the HRF still behaves as a linear function of the physical effort is called the point of deflection (Gellish et al., 2007). This is a good approximation for the threshold HRF, which sports physiologists use to plan the best training program. Measuring the threshold HRF takes place indirectly by analysing a graphical representation of the data.

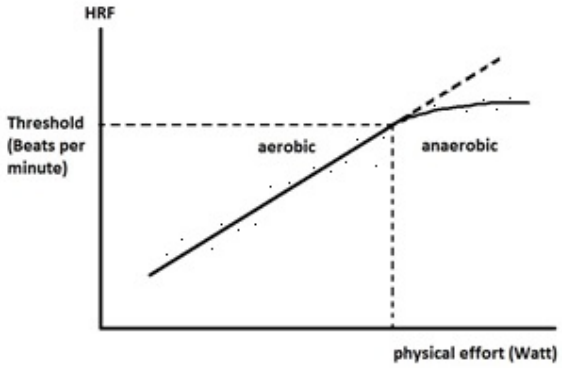

Figure 4. Strategy to find the heart rate threshold (based on Conconi et al., 1987)

During these measurement activities, students had to collect data and reason with variability and make tables and graphical representations of the data using scatter plots, which had not been previously discussed in the school curriculum. In contrast to MA1, the students were asked to represent the measurement data generated in MA2 themselves using Excel^{TM}. For this representation, it was important that students reasoned with variability by considering the particular way that their group performed the test, the measurement devices they used to collect the data (there were several different devices and some measured by hand), and measurement errors. We also expected that students would use investigative strategies in MA2 for finding a trend in their data by using regression lines that they created in Excel.

3.4. DATA COLLECTION AND ANALYSIS

To assess students' reasoning with variability, class discussions and their collaborative work in class were observed and recorded (audio and video) by the first author. Video-

recorded lessons (six lessons of fifty minutes each) were fully transcribed. Data included students' written work, transcripts from video recorded lessons and field notes. One student was not included in the analyses because of her absence during some lessons.

Given our research question on how students can consider variability, we developed an analysis framework (Table 1) to identify which considerations were at stake in the interactions between students or between student(s) and their teacher. This framework is based on Wild and Pfannkuch's (1999) original framework and on Reading and Shaughnessy (2004) who added two more items for educational contexts: describing and representing variability. We decided to add "describing" to noticing and acknowledging variability (NAD) and "representing" to investigative strategy (INV). Using this adapted framework, we analysed students' spoken interactions. Furthermore, we expected the students to demonstrate the four ways in which statistical experts consider variability as described in Table 1.

Table 1. Analysis framework to identify the ways in which students reason with variability

Code	Consideration of variability	Example
NAD	*Noticing, acknowledging and describing variability*	The student implicitly refers to variability, acknowledges it, or explicitly describes variability. For example, when a student mentions that there will be a difference in the results when repeating an experiment.
MEA	*Measuring variability*	The student measures variability in relation to the regression line, for example, calculating the correlation between heart rate and level of physical effort, or between age and maximum heart rate.
EXP	*Explaining variability*	The student tries to explain variability in data, for example, by indicating that people are different, or that the circumstances are not equal.
INV	*Using investigative strategies to handle variability*	The student discusses what is necessary to describe the variability (investigative strategy) or how to handle variability. For example, she represents such variability in a graph or a table to identify a trend, or identifies conditions on which the strategies can be used.

This categorisation (Table 1) is neither a hierarchy nor a list of exclusive categories. The first category in our analysis framework (NAD) is conditional because acknowledging variability is prerequisite for the other three ways of reasoning with variability. Because we wanted unique codes for each utterance, we used code NAD only if no other code applied. We divided the transcripts of the classroom interaction into utterances (our unit of analysis) in which the researcher recognised one of the four ways to consider variability. This process yielded 82 analysis units (utterances). The utterances were independently coded by the first author and an independent researcher who was not involved in this study, but who is an expert in mathematics education and psychology. The interrater agreement measured with Cohen's kappa coefficient was .66, which Cohen (1960) considers substantial. Differences between the two raters were discussed, and the final agreements are presented in the results section.

To judge whether students reasoned with real and induced variability by themselves we also analysed a written task. We asked the students to write down individually a few important elements that they, as a physiotherapist, should consider when collecting the data through the test. The response of students on this question was coded as a) "induced

variability" when they mentioned methodological influences caused by measurers or devices as, for example, measurement errors, or b) "real variability" when they mentioned biological influences as physiological aspects for real variability. Cohen's kappa coefficient in coding these responses was also substantial (.71).

4. RESULTS

We investigated the students' spoken utterances to identify considerations of variability they demonstrated in their reasoning when they engaged in both AuPP activities (MA1 and MA2). Table 2 demonstrates that the measurement activities based on authentic practices stimulated students to reason with variability in all four ways that were identified in the literature on statistical thinking (Table 1), but that the MEA (measuring variability) way of considering variability was not found very often. In the first set of tasks (MA1), 37 utterances were coded, and in the second set of tasks (MA2), 45 utterances were coded. Each student made at least one statement that we could give one of the four codes. The average number of coded utterances per pupil was 6.8 (SD = 5.9). Furthermore, the table indicates that most students attempted to explain (EXP) or investigate variability (INV), but a few tried to explicitly measure (quantify) it (MEA). The reason that Table 2 contains relatively small numbers is that in cases where the teacher explicitly asked about one of the four ways to reason with variability, the reactions of the students were not included in the analysis, because our focus was to investigate whether students themselves would suggest different ways to consider variability. Recall that we used the NAD code only if no other code applied, whereas for the other codes, the students were also aware of variability. For example, when a student discusses variability in relation to the regression line (MEA) the student must have noticed and acknowledged variability in order to do this. However, in such a case we used only the MEA code. The consequence of this decision is that although NAD has the lowest score (see Table 2), it had in fact a 100% score because MEA, EXP, and INV also imply NAD.

Table 2. Number of spoken student utterances about a way to consider variability (n, %)

	MA1					MA2					
	NAD	MEA	EXP	INV	Total	NAD	MEA	EXP	INV	Total	
Total (*n*)	4	5	16	12	37	7	5	18	15	45	
(%)	11%	14%	43%	32%	100%	16%	11%	40%	33%	100%	Total
#Students	2	4	9	7	(11[a])	3	3	6	6	(10)	(12)

[a]The numbers in the parentheses indicate the number of different students

To give a qualitative illustration of students taking into account the four ways they considered variability we briefly report on their considerations during the first measurement activity (MA1). To set the stage, we first sketch how MA1 was introduced. We wanted to involve students in the measurement activities and to become aware of the presence of variability around the regression line. To achieve this, the teacher introduced the following task at the beginning of MA1, asking students to write down their responses: "Consider how a sports physiologist could support a client in improving her fitness and why it can be useful to measure her heart rate for that." Most students wrote down that the HRF depends on the degree of physical effort. None of the students mentioned anything about variability in their written text. In the next sections we provide examples of how

MA1 helped stimulate students to reason with variability in each of the four ways presented in Table 1. [All student names below are pseudonyms.]

4.1. CONSIDERING VARIABILITY IN MA1

Noticing, acknowledging, and describing variability In this subsection we illustrate how students noticed variability. Sometimes support from the teacher was needed to stimulate them to do so. Students were asked to orally explain the difference between the measured RDI and the fitness of their classmates. Although, in their written work, students reasoned with variability in several ways, they did not mention aspects of variability when discussing the problem orally. The teacher (T) responded to this in the next lesson by leading the following discussion about variability.

T:	Do you expect the same [RDI] values when you run the experiment three times?
Tom:	If in between the experiments you recover completely because you have time enough to take a rest, then the [RDI] values will be the same.
T:	Some [students] say if you rest enough, you will find the same RDI [when repeating the experiment].
Jorr:	Well, not quite exactly the same.
T:	So, you expect something close to it?
Jorr:	Yes, a small deviation. You have to take the mean.
Tom:	That is true.

The teacher asked the questions about the students' authentic measurement practices. Their measurement experience and the knowledge they had developed of the AuPP helped them to answer the teacher's questions. In the observations, all students were aware of some variability in their measurements. Students knew that measurement values need not be exactly the same, but sometimes did not seem to see the need to express this variability. By evoking a cognitive conflict (Watson, 2007), the teacher had an important role in stimulating them to express more precisely the difference between what they said and what they observed. First, Tom expected the same values, then later he agreed with Jorr that he expected a small deviation. We therefore coded Tom's second statement and Jorr's statements as acknowledging variability and trying to describe it (NAD).

We conjectured that students would understand that a regression line (in this case for heart rate versus physical effort) is a simplified representation of a relationship and that they would notice the noise (variability) around a signal (regression line). In the following excerpt, Tom considered variability when he described the relation between heart rate and a person's physiology.

Tom:	The deviations between above and beneath [the regression line] become bigger [bigger residuals].
T:	What does it mean?
Tom:	You can use it [regression line] to make predictions.
T:	And how can you use it?
Tom:	Sports physiologists have devices to measure a person's physiology. When you can consider the heart rate as an indicator, it should not be too high or too low. But it is only a partial indicator to measure the effort to do something.
T:	So, you say: it is an indicator. For what?
Tom:	It is an indicator for the degree of effort you must perform to do something.

In reference to the points above and beneath the regression line, Tom reasoned how to measure physiology and identified heart rate to be a "partial indicator" to estimate "the

degree of effort." He seemed to realise that the heart rate gives an indication about a person's physical fitness and that, in fact, you have to use the mean of repeated measurements. Later, Tom also connected the data world to the real world by saying: "See Bert and Elsa: They had a H3 lower than H1. That makes no sense, does it?" (see Figure 2). The teacher acknowledged the importance of Tom's comments and emphasized that an indicator such as heart rate could be used by a sports physiologist as a simplified measure to assess a client's physical fitness, just as an estimate, and that the results of Bert and Elsa may have been caused by real or induced variability.

Measuring variability MA1 confronted the students with variability according to the RDI formula (Figure 2) when they were asked to measure heart rates and find a RDI value to predict the physical fitness of a "client." To explain the RDI formula and the effect of H1 and H2, the teacher showed one of the students' scatter plots consisting of 9 students' measurements (Figure 5), which they collected themselves (the missing students at that time were not finished with their measurements). He asked about the relation between the resting heart rate (H1; Figure 2) and the heart rate after the knee bending (H2). The variability seemed a reason for Bert and Abel to struggle with the relation between their data and the possible regression line for H1 and H2, because it seemed that they indicated something such as induced variability.

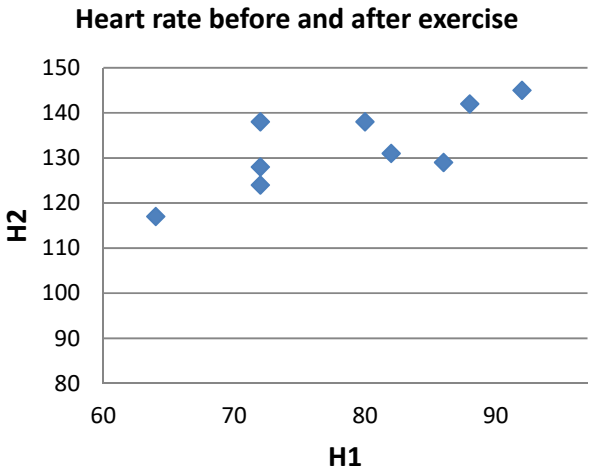

Figure 5. Scatter plot of students' H1 and H2 measurements

Bert:	I can't see a line.
T:	Why?
Bert:	There are only very few points [to recognize a regression line].
Abel:	Actually, I can see a line. But only a little.
Kai:	No.
Abel:	Yes, I do. I see no line, but I recognise a rising regression line. You have to make this data set real.
T:	But this is real data [the students collected it themselves].
Abel:	But I cannot see the shape of the [regression] line.
T:	So, you cannot say something about the relation, but can you say something in common?
Abel:	They [H1 and H2] are linearly proportional.
T:	And how do you know where to draw the regression line?
Abel:	You draw a line with almost the same amount of [data] points above as beneath it.

In explaining the noise as induced variability around the regression line (the signal), Bert realised that a small sample size made it difficult to formulate the relation and that a bigger sample size would be preferable when constructing a meaningful regression line. Executing the experiment and analysing the data probably helped him to have a better understanding of sensible measurements (Enderson, 2003).

Abel recognised a weak pattern in his scatter plot with a low correlation:

T: What value do you expect for the correlation?
Abel: How can you estimate that? They [the points] are very close to each other.
Bert: I think 0.7 or such.
Abel: Yes it must be positive.

At this moment the students had no tools to calculate the correlation. They were asked to estimate the correlation intuitively. They knew about positive and negative correlation and that the value had to be between -1 and 1. Abel expected a clearer trend, "but I cannot see the shape of the line." Thus, despite the small sample and the large variability, Abel was able to recognise a possible linear regression line of the data and realized that there is noise around the regression line by saying that there must be the same number of points above as beneath the regression line.

Explaining variability For the third way of considering variability, we conjectured that they would try to explain variability by indicating that people are different (real variability) or that the circumstances were not equal (induced variability). For example, Rose, who found a higher RDI than she expected, explained: "For me it [RDI value] is not right. I exercise four times a week and swim and cycle every day." When the teacher asked her to further explain this result, Rose shared her heart problems and explained that her heart might not work in a way that is required by the RDI. Other students also considered whether their measurement results fit the formula and tried to explain any deviations from it. Although the students considered themselves physically fit, only three of them found a RDI value that corresponded with their perception of their fitness. When the teacher prompted them to explain these disappointing results, they suggested that the measurements were inaccurate. It seems that the students were urged to express their own motivaton to explain the measurement variability because of their personal knowledge of the authentic context and the disappointing results of the formula (Reading & Shaughnessy, 2004).

They also explained that they made substantial measurement errors because not everyone used the same device to measure the heart rate (induced variability) and that not everyone did the measurement in the same way. Two students used a heart rate monitor, four used a sphygmometer, and the other students did the measurement by hand.

Wim: I think the main difference is caused by different [measurement] devices. We used the sphygmometer which is very good to measure the heart rate. Measuring by hand using a watch can obtain errors.
Mina: Yes, not everyone [fellow students] did the measurements exactly the same. Some bent deeper [in their knee bends] than others and some did the measurements slightly later than the others. Then you get differences.

In addition, students regularly referred to the difference between people to explain the [real] variability (Schwartz, Goldman, Vye, Barron, & Cognition and Technology Group at Vanderbilt, 1998). For example, Bert referred to a famous athlete: "If Louis Armstrong digs his front yard his heartbeat is the same as when an ordinary human is asleep." Bert probably meant cyclist Lance Armstrong, but his message was clear: People are not the

same, or the circumstances are not the same. Students often explained: "A human is not a machine." These utterances represent students' ways to reason with and explain variability.

The task's authentic design encouraged them to notice variability and to investigate links between their data and the scientific context, for example, when they explained why heart rates did not fit the fitness predictions provided by the RDI.

Bert:	Mina, the [RDI] formula implicated that you have a bad physical fitness. Do you smoke?
Mina:	No, I am just not very sporty.
Jorr:	Whether someone smokes cannot be found in the formula, but it [smoking] can have an impact on the [RDI] value.

This transcript is an example of how the authentic character of the measurement activity played a role in students' reasoning with variability. Bert, who saw the RDI as indicative for physical fitness, searched for a contextual explanation for why Mina had an extremely poor fitness level indicated by the RDI value. He was motivated to know more information in order to describe or explain the variability.

Students referred to real and induced variability implicitly probably because they had no prior formal experience with statistical reasoning. When prompted to write down a few important elements that they, as a physiotherapist, should consider when collecting the data for this test, 64 responses were received (each student was allowed to mention more than one element). About two thirds (40) of the responses referred to variability and methodological aspects of induced variability, such as how to do the test. About a third (20) of the responses refered to biological elements (real variability) like the client's physical condition and behaviour before and during the test. Some responses (4) were not about variability.

Using investigative strategies to handle variability The fourth code (INV, see Table 1) was used when the students discussed the requirements and methods to investigate variability. Students represented variability by a scatter plot and considered what was needed to find the regression line. When the teacher asked them what they meant by a rising regression line, students' mathematical knowledge helped them to reason with the variability in their results.

Bert:	We have very few points to draw a regression line.
T:	That is true.
Abel:	[However] I can recognize a [regression] line.
T:	How?
Kai:	The regression line is the mean of all data points.
Abel:	The mean of all data points through the data points.
Kai:	With the same number of points above the line as under the line.
Abel:	It is necessary to have the same number of points above the line so that the overall result of deviations on the upper side is as large as the overall result at the bottom.

Bert was aware that more data points were needed to find a reliable regression line. Abel tried to formulate a version of the "sum of residuals" when mentioning overall results of deviations. In this context, the sum of residuals is the summation of all absolute deviations of the heart rate observations from the regression line. Measuring the data themselves and representing them by a graph seemed to encourage the students to consider the deviations of their measurements from the RDI formula even though they had not learned this idea yet.

4.2. CONSIDERING VARIABILITY IN MA2

There are several differences between MA1 and MA2. Whereas in MA1 students were given the RDI formula, in MA2 they had to investigate their measurements and were informed that the point of deflection in a scatter plot indicates the threshold heart rate. RDI is a simple indicator of physical fitness, based on data from many people. MA2 was more explicitly linked to the biochemical process of metabolic acidosis. This meant that in MA1, students had to reason with variability with regard to their individual values in relation to an aggregate data set, whereas they could remain focused on an individual's data in MA2 when doing their running test (Conconi et al., 1987). In this running test they gathered data by measuring the heart rates with increasing speed of the treadmill. Most students were able to find their own threshold point, but some students did not recognise a trend in their data just as when they struggled with comparing the Ruffier-Dickson formula with the results found by their own data in MA1. Given the literature on students' difficulties of coordinating local and global perspectives on data sets (Ben-Zvi & Arcavi, 2001) and students' limited experience with statistics, this should not be too surprising. What might have helped here is that students had an idea of what the underlying scientific phenomenon was—as in the approach of Lehrer, Kim, and Schauble (2007) in which students had a sense of the true value they were approximating.

Noticing, acknowledging, and describing variability When the students drew their regression line through the points before the the threshold point and the teacher mentioned that this part of their graphs was not totally linear, Alan responded and implicitly mentioned variability:

Alan:	It is not that our heart rate is not linear, but the line is based on something we want to be linear. In my head it is correct. It is not that the heart rate is linear, but because we constructed a linear line as a kind of "guideline."
T:	So you say that, based on this theory, these representations, there must be a linear relation because we think that it is linear?
Alan:	Yes, we invented a linear relation with values that are not linear.
T:	You say: there is no linear relation?
Alan:	No, it is not completely linear. It is almost linear.

In fact, Alan claimed that the linearity of the regression line is invented by people, but that the actual data do not fit the line completely because of variability. Despite that, he did not mention variability explicitly, we can see this excerpt as an example of noticing variability, especially because he mentioned "almost linear" (presumably referring to the strength of the association more than the form).

Measuring variability In the previous section Alan stressed the presence of variability and was not the only student who did this. Later, when they constructed a representation of their collected data in a scatter plot, they noticed variability and named it "a margin." Jorr used the standard deviation of the residuals as a measure for the variability (margin):

Jorr:	You can see this as a "margin."
T:	What does this margin say to you?
Jorr:	The possible deviation for people who score poorly and those who score better.
Elsa:	Yes, a margin!
T:	What is the width of your margin?
Jorr:	2 or 3 times the standard deviation of the deviations from the regression line.

Jorr used the phrase “margin” and Elsa agreed with him. She suggested not sticking rigidly to a formula when advising a client, but to deal with the variability and use a margin. In fact, they used a margin to measure the variability. Jorr mentioned “2 or 3 times the standard deviations of the deviations [residuals],” because he remembered a property of the normal distribution.

Explaining variability In the next excerpt the students tried to explain the variability between heart rate and speed:

Elsa:	It has an aspect of randomness. Like a thermostat. You got a standard, but the value can be below or above. There is a margin of error.
T:	Error?
Elsa:	A range of errors. When the value can be found between two limits and between these limits it is random.
T:	Why errors?
Kai:	I interpret them as errors. Like a standard. It can be above or below. It fluctuates between them. The processes in your body are never the same.
Kai:	Your body is not a machine.
T:	What do you mean by that?
Kai:	I mean, your body is not always working as described by the formula.
Alan:	Suppose that during training you see a beautiful woman, then your heart rate will become higher.
Elsa:	The treadmill is rather long. So, you can make big steps or small steps. Or walk a little faster and slow down a little.

Again the students explained the real variability (e.g., “your body is not a machine,” “you can make big steps or small steps”). The students agreed that even if the tests were done in a laboratory, the results would still vary.

Using investigative strategies to handle variability To illustrate that the students were able to find the threshold point and handle variability, we present another example. Investigating the collected data was not easy for every student, because some outliers were found. Despite those, most students were able to find a trend using a regression line. Despite an outlier (Figure 6, fifth dot from left), Derk was able to recognise a linear trend for the first 12 measurements. He drew a straight line by eye to fit the first 12 points and based on this line, he expected his threshold point between the twelfth and thirteenth measurement result to be about 16 km/h.

Despite that some students had to deal with more variability, most students were able notice this variability and to investigate it in order to find the threshold point. Only a few with very scattered data struggled with finding the threshold point. This research showed indications that the authentic character of this task can help students to investigate their measurement results and analyze them with the help of regression techniques in order to find an answer for a “real” problem.

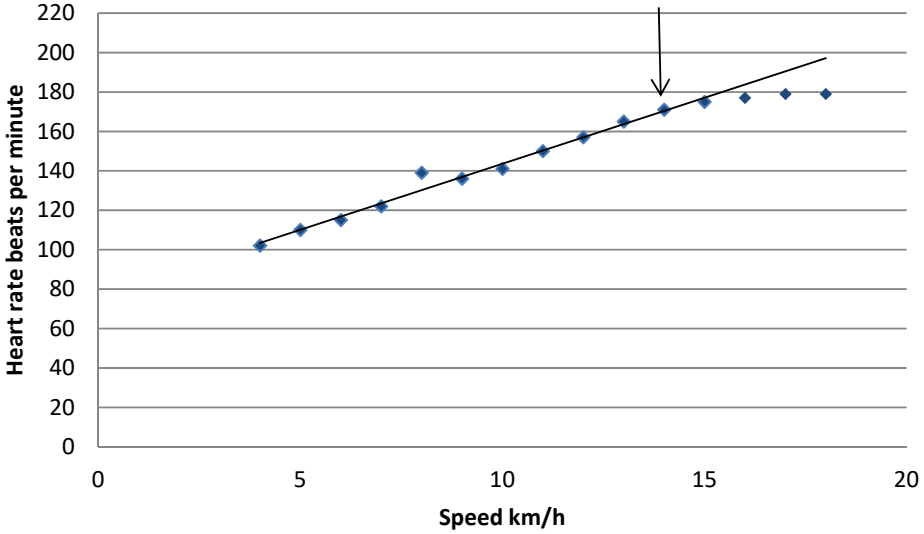

Figure 6. Derk's scatter plot of his heart rates at increasing speeds with a straight line fit by eye and an arrow indicating the point of deflection

5. DISCUSSION AND RECOMMENDATIONS

Variability is a key concept in statistics that typically does not receive the attention it deserves in school statistics. As part of a larger research project we designed measurement tasks to promote meaningful reasoning with variability. These tasks were based on AuPPs in which regression was used. The advantage of measurement is that it is at the interface between context and statistics, where students can get a feel for where variability comes from (e.g., variability in the phenomenon studied versus measurement error). The assumed advantage of basing measurement tasks on an AuPP is that students may see the need for learning about statistical techniques such as regression with taking variability into account and thus be motivated to learn about them. Moreover, it is known that students often demonstrate computational habits without realistic considerations when they solve word problems (Cooper & Harries, 2002). More recent studies indicate that more authentic tasks can help to counteract such habits (Verschaffel, Greer, Van Dooren, & Mukhopadhyay, 2009).

To investigate whether our measurement tasks did help students to reason with variability in rich ways, this article addressed the following research question: *In what ways do secondary students consider variability within measurement activities based on authentic professional practices?* We used an analysis framework based on ways in which statisticians consider variability (Table 1) to analyse how students reasoned with variability: noticing, acknowledging, and describing variability; measuring variability; explaining variability; and using investigative strategies to handle variability. We found that students demonstrated all these ways of considering variability (though of course in less advanced ways than would statisticians), but that measuring variability did not occur as often as explaining and investigating variability. More research is needed to understand how to design instruction to promote this particular way of considering variability. We know that students' consideration of variability is generally very poor. Our research gives a proof of principle that instruction can be designed to prompt students to consider

variation, and we hope it inspires others to try similar ideas so that the transferability of the ideas can become clear.

In more detail, all students noticed and acknowledged variability. They experienced that the data they found did not exactly fit the RDI formula (MA1) or regression line between the resting heart rate and the heart rate after the knee bending (MA2), and they tried to find explanations for this. Furthermore, they were concerned with a "margin" (their expression for variability) around the graph of the RDI formula to predict a client's fitness, and they explained the deviation of their own fitness values to the value obtained by the formula they found. To control the variability in data the students suggested taking more measurements or using the same device for every measurement. To identify a suitable sports program the students used investigative strategies to find the threshold point. Finally, we conclude that the two activities supported students' reasoning with the four ways to consider variability as described in our coding scheme, but as mentioned before, "measuring variability" (MEA) was only found in a small portion of the utterances. In this study we had no means to explain this and therefore suggest to further examine this issue in future research.

Because of the deviations from their own data and the data predicted by the RDI, students were asking themselves whether heart rate is the key characteristic that is needed to measure physical fitness. They suggested that not everyone did the measurement in the same way, but that the heart rate can be indicative for this context about physical fitness. The "noise" in their data urged them to consider the sources of variability (Konold & Pollatsek, 2002). For example, some students noticed that other students did not apply the same methods. The findings suggest that MA1 and MA2, as examples of measurement tasks based on the AuPP of a sports physiologist, stimulated students to consider variability in all these ways, provided that the teacher helped them to deepen their consideration of variability.

As the analysis suggests, performing measurements within authentic practices seemed to stimulate students to reason with variability in different relevant ways. We suggest that the students gained an understanding of "sensible" measurements by using measurement and investigative activities to find patterns, such as a trend. They found trends by representing, analysing, and generalising their collected data in table formats as well as in graphs. Their prior contextual knowledge helped them to acknowledge and deal with variability (e.g., "your body is not a machine"), but the teacher's support was often needed to elicit students' consideration of variability. The transcripts suggest that students were not just solving a word problem, but considered variability to find an answer for a "real" problem. For most students, these activities based on authentic practices were successful in promoting students to reason about variability in ways that we envisioned. Some students struggled with explaining variability in their measurement results, but our study suggests that the students gained awareness that you could use RDI for sensible predictions, even though it does not precisely describe reality. The measurement experiences of the students, together with the class discussions, contributed to the students' view that RDI is just a simplification of reality. For some students, this was difficult to see initially and some struggled with the variability of the data. However, the results of this study suggest that the activities generally supported students in learning to measure parameters of physical fitness using investigative techniques and to reason with variability in valuable ways. We presume that other AuPPs, such as measuring the correctness of thermometers in order to calibrate them, could give similar results.

As a limitation of our study, we note that only one small group of students was involved. We consider this study a proof of principle that it is promising to base tasks in statistics education on AuPPs in which statistical techniques are used. However, scaling up

to larger groups, more schools, and more teachers is needed. It is also recommended to perform a comparative study by comparing groups that may or may not carry out authentic measurements. Furthermore, we think that the measurement activities can be extended to support students in understanding other more sophisticated types of variability as well, such as sampling variability (e.g., Ben-Zvi, Bakker, & Makar, 2015; Pfannkuch, Arnold, & Wild, 2015).

In follow-up research (Dierdorp et al., 2014), we found that basing teaching and learning strategies on AuPPs may also help students to be motivated to learn and see the point of their learning and to help to develop rich multifaceted concepts. We suggest that our strategy, based on AuPPs in which statistics is used, is a promising direction of research and design that can help students in making connections between school subjects such as mathematics and science.

REFERENCES

Adams, T. L., & Harrell, G. (2003). Estimation at work. In D. H. Clements & G. Bright (Eds.), *Learning and teaching measurement* (pp. 229–244). Reston, VA: National Council of Teachers of Mathematics.

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

Barab, S. A., & Squire, K. D. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63.
[Online: http://iase-web.org/documents/SERJ/SERJ3(2)_BenZvi.pdf]

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*, 35–65.

Ben-Zvi, D., & Aridor, K. (2016). Children's wonder how to wander between data and context. In D. Ben-Zvi & K. Makar (Eds.), *Teaching and learning of statistics: International perspectives* (pp. 25–36). Chaim, Switzerland: Springer International Publishing Switzerland.

Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics*, *88*(3), 291–303.

Buys, K., & de Moor, E. (2005). Domain description measurement. In M. van den Heuvel-Panhuizen & K. Buys (Eds.), *Young children learn measurement and geometry* (pp. 15–36). Utrecht, The Netherlands: Freudenthal Institute (FIsme).

Cobb, G. W., & Moore, D. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Conconi, F., Ferrari, M., Ziglio, G., Droghetti, P., & Codeca, L. (1987). Determination of the anaerobic threshold by a noninvasive field test in runners. *Journal of Applied Physiology*, *52*(4), 869–873.

Cooper, B., & Harries, T. (2002). Children's responses to contrasting "realistic" mathematics problems: Just how realistic are children ready to be? *Educational Studies in Mathematics*, *49*(1), 1–23.

delMas, R. C., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82.
[Online: http://iase-web.org/documents/SERJ/SERJ4(1)_delMas_Liu.pdf]

Dewey, J. (1929). *The quest for certainty: A study of the relation of knowledge and action.* New York: Minton, Balch, & Company.

Dierdorp, A. (2013). *Learning correlation and regression within authentic contexts.* Utrecht, The Netherlands: CD Bèta Press.

Dierdorp, A., Bakker, A., Eijkelhof, H. M. C., & van Maanen, J. A. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning*, *13*(1-2), 132–151.

Dierdorp, A., Bakker, A., van Maanen, J. A., & Eijkelhof, H. M. C. (2014). Meaningful statistics in professional practices as a bridge between mathematics and science: An evaluation of a design project. *International Journal of STEM Education, 1*(9), 1–15. doi: 10.1186/s40594-014-0009-1

Eijkelhof, H. M. C., & Krüger, J. (2009, September). *Improving the quality of innovative science teaching materials.* Paper presented at European Science Education Research Association conference (ESERA), Istanbul, Turkey.

Enderson, M. (2003). Using measurement to develop mathematical reasoning at the middle and high school levels. In D. H. Clements & G. Bright (Eds.), *Learning and teaching measurement* (pp. 271–281). Reston, VA: National Council of Teachers of Mathematics.

English, L. D. (2009). Promoting interdisciplinarity through mathematical modeling. *ZDM - The International Journal on Mathematics Education*, *41*(1), 161–181.

Estepa, A., & Sanchez-Cobo, F. T. (1998). Correlation and regression in secondary school textbooks. In L. Pereira Mendoza, L. Seu, T. Wee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on the Teaching of Statistics* (Vol. 2, pp. 671–676). Voorburg, the Netherlands: International Statistical Institute.

Fitzallen, N. (2012). Interpreting graphs: Students developing an understanding of covariation. In J. Dindyal, L. P. Cheng, & S. F. Ng (Eds.), *Mathematics education: Expanding horizons (Proceedings of the 35th annual conference of the Mathematics Education Research Group of Australasia)*. Singapore: MERGA.
[Online: https://www.merga.net.au/documents/Fitzallen_2012_MERGA_35.pdf]

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., Scheaffer, R. (2005). *Guideslines for Assessment and Instruction in Statistics Education (GAISE) Report: A pre-K-12 curriculum framework.* Alexandria, VA: American Statistical Association.
[Online: http://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12_Full.pdf]

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*(1), 92–99.
[Online: http://iase-web.org/documents/SERJ/SERJ4(1)_Garfield_BenZvi.pdf]

Gellish, R. L., Goslin, B. R., Olson, R. E., McDonald, A., Russi, G. D., & Moudgil, V. K. (2007). Longitudinal modeling of the relationship between age and maximal heart rate. *The American College of Sports Medicine*, *39*(5), 822–829.

Gooya, Z., Khosroshahi, L. G., & Teppo, A. R. (2011). Iranian students' measurement estimation performance involving linear and area attributes of real-world objects. *ZDM – The International Journal on Mathematics Education*, *43*(5), 709–722.

Hjalmarson, M. A., Moore, T. J., & delMas, R. (2011). Statistical analysis when the data is an image: Eliciting student thinking about sampling and variability. *Statistics Education Research Journal*, *10*(1), 15–35.
[Online: http://iase-web.org/documents/SERJ/SERJ10(1)_Hjalmarson.pdf]

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289.

Lee, H. S., & Butler-Songer, N. (2003). Making authentic science accessible to students. *International Journal of Science Education*, *25*(1), 1–26.

Lehrer, R. (2003). Developing understanding of measurement. In J. Kilpatrick, W. G. Martin, & D. E. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics* (pp. 179–192). Reston, VA: National Council of Teachers of Mathematics.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*(2), 116–133.

Lehrer, R., Kim, M., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in modeling and measuring variability. *International Journal of Computers for Mathematical Learning*, *12*(3), 195–216.

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

MacGillivray, H. (2004). Coherent and purposeful development in statistics across the education spectrum. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable*. Voorburg, The Netherlands: International Statistical Institute.

Makar, K., & Confrey, J. (2007). Moving the context of modeling to the forefront: Preservice teachers' investigations of equity in testing. In W. Blum, P. Galbraith, H-W. Henn, & M. Niss (Eds.), *Modeling and applications in mathematics education* (pp. 485–490). New York: Springer.

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.

Paulet, G., Gratas, A., Dassonville, J., & Rochcongar, P. (1981). Comparative study of four effort tests on nine-year-old children. *European Journal of Applied Physiology*, *46*(1), 55–68.

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach* (1st ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics*, *88*(3), 343–360.

Prins, G. T. (2010). *Teaching and learning of modeling in chemistry education. Authentic practices as contexts for learning.* Utrecht, The Netherlands: CD Bèta Press.

Rabinovich, S. G. (2005). *Measurement errors and uncertainties – theory and practice.* New York: Springer.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201–226). Dordrecht, The Netherlands: Kluwer.

Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal*, *7*(1), 40–59.
[Online: http://iase-web.org/documents/SERJ/SERJ7(1)_Reid_Reading.pdf]

Roth, W. M., & Bowen, G. M. (2003). When are graphs ten thousand words worth? An expert/expert study. *Cognition and Instruction*, *21*(4), 429–473.

Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & Cognition and Technology Group at Vanderbilt (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (1st ed., pp. 233–273). Mahwah, NJ: Lawrence Erlbaum Associates.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F.K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2. pp. 957–1010). Charlotte, NC: Information Age.

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

Smith III, J. P., van den Heuvel-Panhuizen, M., & Teppo, A. R. (Eds.) (2011). Learning, teaching, and using measurement: Introduction to the issue. *ZDM – The International Journal on Mathematics Education*, *43*(5), 617–610.

Sorto, M. A. (2006, July). Identifying content knowledge for teaching statistics. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil [CDROM]. Voorburg, the Netherlands: International Statistical Institute.

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

Verschaffel, L., Greer, B., Van Dooren, W., & Mukhopadhyay, S. (Eds.). (2009). *Words and worlds: Modeling verbal descriptions of situations.* Rotterdam/Boston/Taipei: Sense Publishers.

Watson, J. M. (2007). The role of cognitive conflict in developing students' understanding of average. *Educational Studies in Mathematics*, *65*(1), 21–47.

Westbroek, H. B., Klaassen, C. W. J. M., Bulte, A., & Pilot, A. (2010). Providing students with a sense of purpose by adapting a professional practice. *International Journal of Science Education*, *32*(5), 603–627.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–265.

ARTHUR BAKKER
Utrecht University
Freudenthal Institute
Princetonplein 5
3584 CC Utrecht
The Netherlands