# What is Hidden Beyond the Data[1]?
# Young Students Reason and Argue about Some Wider Universe[2]

Working Version

*Dani Ben-Zvi*, *Einat Gil*, and *Naomi Apel*

University of Haifa

## 1. Overview

We are developing an epistemological/conceptual framework for *Informal Inferential Reasoning* (IIR), which includes both cognitive and socio-cultural aspects of learning with an emphasis on argumentation. This theoretical framework supports (and at the same time evolves by) the design of a socio-constructivist open-ended Exploratory Data Analysis (EDA) learning environment facilitated by *TinkerPlots*, and the longitudinal study of students' learning in this environment. In this ongoing three-year development and research project (The *Connections* Project, grades 4–6, 2005–2007) we focus on the study of students' emerging statistical reasoning and argumentation skills within an empirical statistical enquiry cycle. The instructional and research activities of the Project's first two years (grades 4 and 5) focused on reasoning about distribution, variability and comparing groups and aimed at preparing the students for learning IIR in the third year (2007, details provided below). In the current study, we follow one sixth grade class (age 11–12, n=25) during their third year in the *Connections* Project. We try to outline students' learning processes and development of IIR and argumentative skills with a focus on informal reasoning about sampling and parameter estimation by conducting surveys and drawing simple random samples. In the SRTL-5 Forum, a preliminary qualitative analysis of carefully selected four video vignettes and other supporting data will be presented and a description of what it may mean to begin reasoning and arguing informally about statistical inference by young students will be proposed.

## 2. Theoretical Background
## Informal Inferential Reasoning (IIR) and Argumentation

*Statistical inference* is a theory, standard method and practice that attempts to draw a conclusion about a particular population from data-based evidence provided by a random sample. Statistical inference "moves beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are uncertain" (Moore, 2004, p. xxiv). There are two important themes in statistical inference: a) *parameter estimation* – generalizing from a small sample to a larger population by conducting

---

[1] EDA activities are often schematized by the slogans - looking at the data (preliminary analysis), looking between the data (comparisons), **looking beyond the data (informal inference)** and looking behind the data (context) (Curcio, 1989; Shaughnessy et al., 1996).

[2] Ben-Zvi, D., Gil, E., & Apel, N. (August, 2007). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), Reasoning about Informal Inferential Statistical Reasoning: A collection of current research studies. *Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy* (SRTL-5). University of Warwick, UK, August 11-17, 2007.

surveys and using confidence intervals; and b) *hypothesis testing* – determining if a pattern in the data is due to cause and effect by conducting experiments and using significance tests.

Reasoning about *data analysis* and reasoning about *statistical inference* are essential for effective work with data and gaining understanding from data. We suggest positioning *Informal Inference* as the "bridge" between exploratory data analysis (EDA) and formal statistical inference. While the purpose of EDA is unrestricted exploration of the data, searching for interesting patterns, the purpose of statistical inference is to answer specific questions, typically posed before the sample data are produced. Conclusions in EDA are informal, inferred based on what we see in the data, and apply only to the individuals and circumstances for which we have data in hand. In contrast, conclusions in statistical inference are formal, backed by a statement of our confidence in them, and apply to a larger group of individuals or a broader class of circumstances. We now suggest a preliminary working "definition" of IIR:

> *Informal Inferential Reasoning* (IIR) refers to the cognitive activities involved in informally drawing conclusions or making predictions about "some wider universe" from patterns, representations, statistical measures and statistical models of random samples, while attending to the strength and limitations of the sampling and the drawn inferences.

Underlying the development of IIR, as defined above, are fundamental statistical thinking elements: consideration of variation (Wild & Pfannkuch, 1999) and reasoning about distribution, center, comparing groups and sampling within an empirical enquiry cycle (Pfannkuch, 2006). IIR is conceptualized similarly by Rubin, Hammerman and Konold (2006) as statistical reasoning that involves consideration of multiple dimensions: properties of data aggregates, the idea of signal and noise, various forms of variability, ideas about sample size and the sampling procedure, representativeness, controlling for bias, and tendency. A theoretical framework of inference is suggested by Bakker, Derry and Konold (2006) that broadens the meaning of statistical inference to allow more informal ways of reasoning and to include human judgment based on contextual knowledge.

The *argumentation* metaphor is used by Ben-Zvi (2006) to emphasize the role of argumentative elements in more informal ways of statistical reasoning in the context of primary school students using *TinkerPlots*. He points out that informal inference is closely related to argumentation: Deriving logical conclusions from data-based evidence - whether formally or informally - is accompanied by the need to provide persuasive arguments based on data analysis. Argumentation refers to discourse for persuasion, logical proof, and evidence-based belief, and more generally, discussion in which disagreements and reasoning are presented (Kirschner, Buckingham Shum, & Carr, 2003).

We suggest that integration and cultivation of both informal inference and informal argumentation are essential in constructing students' statistical knowledge and reasoning in rich learning contexts. Argumentative activities were found beneficial for knowledge building and evaluation of information in some conditions (Schwarz, Neuman, Gil, & Ilya, 2003). This view is supported by Abelson (1995), who proposes two essential dimensions to informal argumentation: The act or process of deriving conclusions from data (inference), and providing persuasive arguments based on the data analysis (rhetoric and narrative). The use of

the argumentation metaphor for IIR means accounting for the following elements: Applying the language of arguing about a statistical claim whether a claim is believed to be true, the level of confidence in the claim trueness, the role of data-based evidence and using that evidence well, what it takes to be convinced that the claim is true or false, and the limitation and application of a claim.

The emerging vision of IIR that we continue to develop during the current study combines cognitive aspects of reasoning about data and chance as well as socio-cultural aspects embedded in classroom and individual practices, dispositions and discourse (see Figure 1). The purpose of this theoretical framework is to support (and at the same time to be influenced by) the design of learning, teaching and curriculum, and the study of multiple perspectives of developing students' IIR as a bridge between exploratory data analysis and statistical inference.

**Informal Inferential Reasoning (IIR)**

| Cognitive Aspects | Socio-Cultural Aspects |
|---|---|
| • Reasoning about Variability<br>spread, density, variability from a variety of sources, … | • Instructional Context<br>Learning environment design, teachers and students' awareness of purposes and utility,… |
| • Distributional Reasoning<br>Aggregate views, pattern and trend, hypothesis and prediction, as well as local reasoning about individual cases, outliers,… | • Language<br>Discourse types and norms to discuss data, graphs, sampling, inferences, … |
| • Reasoning about Signal and Noise<br>Center, measures, modal clumps, summary,... | • Culture and History<br>Student's beliefs, dispositions, prior knowledge and background, ... |
| • Sampling Reasoning<br>Sample size, randomness, sampling variability and behavior, bias, representativeness, ... | • Argumentation<br>Arguing about inferences, claims and counterclaims, data-based evidence, … |
| • Contextual Reasoning<br>Interpretation, alternative explanations, … | • Socio-Statistical Norms<br>classroom discourse norms, what a statistical claim is, what it takes to be convinced that a claim is true or false, … |
| • Graph Comprehension<br>Creating and decoding visual shapes, … | • Evaluative Disposition<br>Providing and assessing evidence, level of confidence, critical disposition to sampling and inference.… |
| • Reasoning about Comparing Groups<br>Comparison of center, spread and shape, … | • Flexibility<br>Transfer back and forth between local and global view of data, sample and population, data and context, reality and its representations, … |
| • Probabilistic Reasoning<br>Uncertainty, random events, chance, … | |
| • Inferential Reasoning<br>Generalizations, limitations and strength of conclusions, … | |

*Figure 1*: Suggested theoretical framework of IIR.

## 3. The Study

These ideas and framework formed the motivation to explore the possibilities for senior students in primary school (grade 6) – with some prior statistical knowledge (most of them

participated in the previous two years of the *Connections* Project) – to develop an informal understanding of statistical sampling and inference in argumentative rich contexts using *TinkerPlots*. This learning environment involves authentic, rich and open-ended EDA investigations that are iteratively designed before and during the research period. These activities include peer collaboration and group discussions, whole class guided argumentative discourse, organizing and graphing data and reasoning about data with *TinkerPlots*, and guidance by math teachers and members of the research team. In the following sections, a brief and essential background of students' learning "history" in the *Connections* Project is provided, followed by the current research questions and methods.

### 3.1 The *Connections* Learning Environment Design

The *Connections* Project is a three-year (2005–2007) research and development project at grades 4–6 (10–12 year-olds) that takes place in a science-focused magnet primary school in Haifa. The investigators, mathematics educators and statistics education researchers from the University of Haifa, worked with primary school teachers and students to study students' evolving ideas of statistical reasoning in a computerized learning environment. Students actively experienced some of the processes involved in experts' practice of data-based enquiry by working on small data scenarios, investigated by peer collaboration and classroom discussions. In *Connections* Students work on two sequential strands - (1) classroom data investigations activities, and (2) a 'research project' that is an extended activity, in which the students act as independent and responsible learners. Students generate and phrase the questions they wish to investigate, suggest hypotheses, analyze data, interpret the results and draw conclusions. At the end they present their main results to fellow students and parents in a 'statistical happening'.

A central feature of the *Connections* Project is the use and the study of *TinkerPlots* (Konold & Miller, 2005), a statistical visualization tool that is designed to help students develop statistical reasoning and learn new ways of representing data. Students can begin using *TinkerPlots* without knowledge of conventional graphs or different data types, without thinking in terms of variables or axes. By progressively organizing their data (ordering, stacking, and separating data icons), students gradually organize data to answer their questions and actually design their own graphs. In fact, our observations show that *TinkerPlots* becomes a thinking tool for these students, namely, it efficiently scaffolds their ongoing negotiations with data, statistical ideas and inferences.

| Year | Fourth Grade | Fifth Grade | Sixth Grade |
|---|---|---|---|
| **2005** | Data, distribution, statistical inquiry, basic comparing groups, basic *TinkerPlots* skills<br>- Three grades | | |
| **2006** | Same as above, but with less involvement of the research team.<br>- Three grades | Comparing groups, variability and center, samples (growing samples), basic argumentative skills, advanced *TinkerPlots* skills<br>- Three grades | |

| | | | |
|---|---|---|---|
| **2007** | Same as above, but with less involvement of the research team.<br>- Two grades | Same as above, but with less involvement of the research team.<br>- Three grades | IIR, sampling (e.g., SRS, bias, connection to inference), advanced argumentative skills, advanced *TinkerPlots* skills<br>- Three grades |

*Figure 2*: The *Connections* evolving themes and the number of grades (about 25 students in any grade).

*The Connections* Project was planned as a growing and expanding initiative. We started with fourth graders and added one more grade every subsequent year. Figure 2 presents the evolving statistical themes of the Project for each grade level as well as the number of grades that took part in the study at each stage. The dark grey boxes are the "focus grades" at each year, e.g., in 2006 the research team designed the instructional materials for fifth grade and studied the learning of three fifth grade classes (about 75 students). In 2007, mathematics teachers used improved versions of these materials in their fifth grade with less direct involvement of the research team than in 2006 (these are represented by lighter grey boxes in Figure 2).

### 3.2 Year 1, Fourth Grade (2005)
### Reasoning about Data and Distribution

In 2005 students were first introduced to the basic statistical ideas (e.g., data, distribution) and to the *TinkerPlots* software in the context of experiencing several times the empirical enquiry cycle. The initial research focused on students' prior knowledge and statistical reasoning about data and distributions, design principles of the learning environment, and the role of the technological tool in assisting the development of student's statistical reasoning (Ben-Moshe, 2007).

A typical fourth grade activity is the "Backpack Activity" (one of the demo activities in *TinkerPlots*), in which teams of students examine student backpack weights in relation to student body weights. They are introduced to expert recommendation to carry a backpack not heavier than 15% of body weight, and that heavy loads can cause shoulder pain or lower-back pain. They are given authentic data of four different grades in their school (part of which they collect on their own): one, three, five and seven and calculate the percentages by dividing the weight of a student's backpack by the student's weight to compare what students were carrying with doctor recommendations. Ben-Moshe (2007) reports that students became fluent in using basic statistical concepts and skills, well versed with the technological tool, and sensibly interpreting some of the graphs that they constructed. Typical ways of students' distributional reasoning using data grouping techniques and interpretation are identified.

### 3.3 Year 2, Fifth Grade (2006)
### Reasoning about Center, Variability and Comparing Groups

In the Project's second year, 2006, we continued working with the same cohort of students to study their evolving ideas of center and variability, comparing groups, and argumentative skills. Math and science teachers collaborated in guiding their students to actively model and analyze natural, sometimes complex, systems (for example, air pollution, water consumption)

using statistical descriptions. The mathematics teachers guided their students through a series of genuine mini exploratory data analysis projects while the science teachers provided the scientific background and inquiry skills for the main research project (see Figure 3 for the main components of this year's Project).

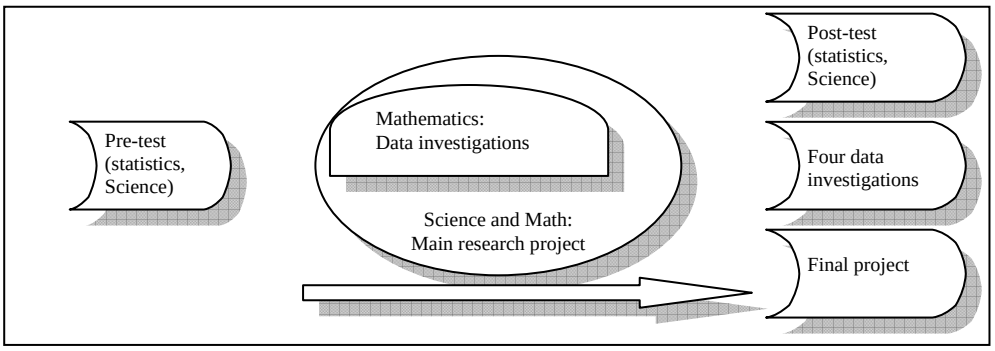


*Figure 3*: Outline of the fifth grade *Connections* 2006 learning environment.

Fifth grade students collected and investigated real data about themselves and peer students and compared them to sample data generated from the UK *CensusAtSchool* data base (http://www.censusatschool.ntu.ac.uk/default.asp). Students were gradually introduced to increasing sample sizes: from one case (the student herself), to a very small sample of four familiar cases, to about 8, 16, 80, and 240 cases (each small sample is a subset of the larger samples). Following the growing samples instructional heuristic (Bakker & Gravemeijer, 2004), we encourage fifth grade students to reason with stable features of variable processes, and compare their hypotheses regarding larger samples with observations generated by them from real data. This process not only helped students get a good grip of the data at hand, but also supported their statistical reasoning by observing aggregate features of distributions, identifying signals out of noise, accounting for the constraints of their inferences, and providing persuasive data-based arguments.

In order to collect real and relevant data, a 19-item questionnaire about gender and age, body measurements, home to school distance and time, computer and cellular phone ownership, etc. was used. Each fifth grader was assigned to collect data from randomly selected three students in grades 2, 4 and 6 as well as herself, and enter it to a *TinkerPlots* file. At this stage the class discussed issues of posing questions, census and measurement.

Grouped in couples, students investigated first a small sample (n=8, four cases of each one of them), formulating questions they find interesting (for example, studying the association between height and arm span), proposing hypotheses, and continuously testing them by constructing and interpreting plots in *TinkerPlots*. Having generated several data-based interpretations, students were asked to hypothesize whether their conclusions would hold in larger samples. Students were then grouped in quartets to test their hypotheses and discuss, confirm or refute their conclusions based on a larger sample of about 16 cases (eight from each couple). The whole class discussed later the quartets' inferences using an overhead projector. Energized by teacher and peers' observations and fresh ideas, the original couples depart back to the computer lab to further investigate their previous conjectures as well as new ideas, using this time a larger sample – about 80 students – generated by the whole class.

This cycle of data generation and analysis, data-based interpretation, class presentation and reflection repeated twice on larger samples. At the final stage, students compared a sample of about 240 students from their own school with 200 UK students (see an outline of the various stages in Figure 4).

At the same time, students worked in the science classroom on their main research project, analyzing natural systems using statistical descriptions. In a festive event, students presented and discussed their main research project in front of their peers, teachers and parents.

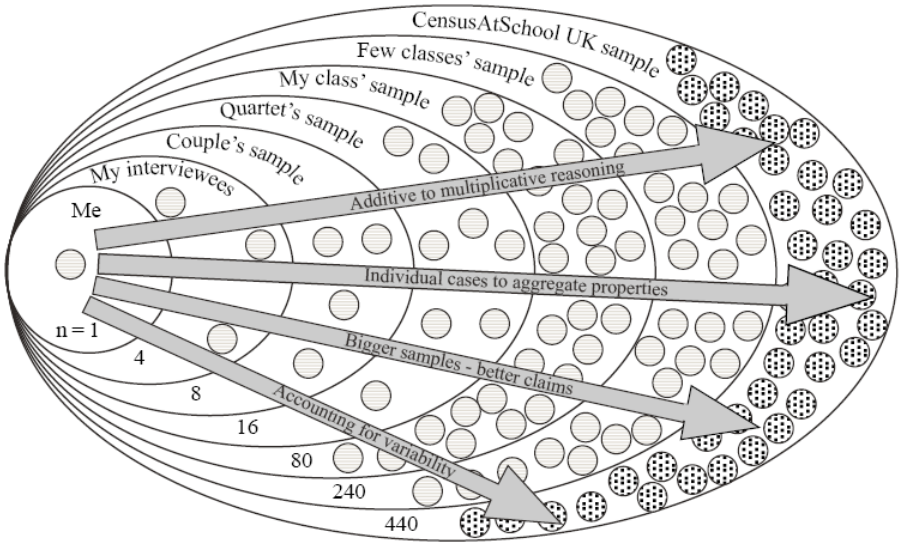


*Figure 4*: The growing sample sequence and the key changes in students' statistical reasoning.

Initial results (Ben-Zvi, 2006) show that the careful design of this learning trajectory and the classroom argumentative discourse coupled with the unique features of *TinkerPlots* were instrumental in supporting students' multiplicative reasoning, aggregate reasoning, acknowledging the value of large samples, and accounting for variability. These processes were accompanied by greater ability to verbalize, explain and argue about data-based informal inferences. These were all taken as emerging statistical reasoning and argumentative skills that were assumed as the sound basis of the following year's experiment (sixth grade, focus on IIR).

It is important to note that the use of samples in fifth grade classes was relatively limited and referred only to the growing samples metaphor. It did not refer to the broad issues of reasoning about samples and sampling that are addressed more broadly at sixth grade as a basis for informal inference. The significant improvement and the high achievements in the administered statistical knowledge tests (described briefly in the next section) encouraged us to design an innovative and challenging IIR learning trajectory for sixth grade (described below).

### 3.4 Growth of Statistical Knowledge

In a longitudinal study, we administered every year (2005-2007) a statistics pre- and post-tests. The fourth and fifth grade test focused on statistics literacy while sixth grade test focused on sampling and IIR. The ten statistics literacy items were taken from the released items of TIMSS (1995). The results of this quantitative analysis were very positive. For example, in order to examine change in students' statistical knowledge across the first two years, we used repeated measure ANOVA procedure for six questions that were identical in the fourth and fifth grades (42 students who participated in all four tests). The descriptive statistics of this analysis (Table 1 and Figure 5) shows continuing improvements in students' results in the statistics literacy tests.

| | **Mean** | **Std. Deviation** | **N** |
|---|---|---|---|
| **Pre 2006, 4th grade** | 6.61 | 2.53 | 42 |
| **Post 2006, 4th grade** | 8.43 | 1.88 | 42 |
| **Pre 2007, 5th grade** | 8.69 | 1.63 | 42 |
| **Post 2007, 5th grade** | 9.05 | 1.32 | 42 |

*Table 1*: Descriptive statistics of the repeated measure procedure.

Significant differences were found between pre 06 and post 06 ($F_{(1,41)}$=23.71, $p$<.0001); post 06 and post 07 ($F_{(1,41)}$=6.18, $p$<.05); pre 06 and pre 07 ($F_{(1,41)}$=29.34, $p$<.0001). Although not a statistically significant difference, we were encouraged by the positive difference between pre 07 (April 2007) and post 06 (April 2006), which was carefully taken as a good retention rate of statistical facts and skills. We also found that the impact of the intervention measured by pre-post average differences was significantly larger ($t_{(41)}$= 4.01, $p$<.0001) in 2006 (M=1.83, Sd=2.43) than in 2007 (M=0.36, Sd=1.24). This can be explained by the much greater involvement of the research team in fifth grade in 2006 than in 2007, when we were focused on the sixth graders IIR intervention. Further details of this and other quantitative analyses will be reported elsewhere. The analysis of 2007 sixth grade IIR knowledge pre- and post- tests is currently underway.

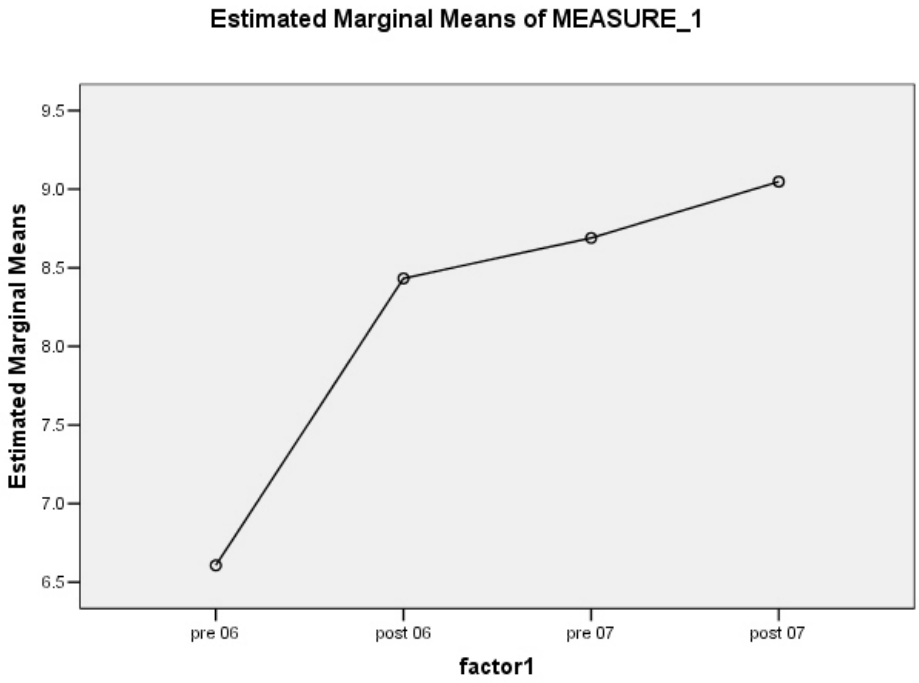


*Figure 5*: Estimated marginal means of students' grades in 2006-7 pre-and post- tests in fourth and fifth grades.

### 3.5 Current study, Sixth Grade (2007)
### IIR and Reasoning about Samples and Sampling

In 2007 (third year of the *Connections* Project, grade 6) we study students' emerging IIR with a focus on ideas of samples and sampling, their relationship to informal statistical inference, and improved argumentative praxis. The designed IIR learning trajectory (Gil, 2007) provides ample opportunities for students throughout the five-week intervention to account for, describe and argue about variability in samples, sampling bias, and representativeness as they make informal inferences about how these samples relate to the population from which they were drawn, and whether these samples lead them to infer claims about what that population might be.

#### 3.5.1 Research Questions

The following research questions are used to structure the current study and the analysis of data collected: How do sixth grade students begin to reason about informal inference in a rich and supportive IIR learning environment? What characterizations of the learning environment and instructional activities support the development of IIR? What are the characterizations of argumentative activities that help develop students' IIR? What sorts of classroom discourse are effective in supporting the development of students' IIR? What is the role of the teacher in modeling IIR and moderating the classroom argumentative discourse about statistical inference? How *TinkerPlots* assists in promoting IIR?

#### 3.5.2 Method

We carried out a developmental research (cf., Ben-Zvi, Garfield and Zieffler, 2006) to investigate students' construction of meanings and knowledge and improve the pre-formulated instructional design by checking and revising conjectures about the trajectory of learning for both the group and the individual students who compose the experiment population. The research was intensive (2-4 meetings a week, for 5 weeks) and somewhat invasive, in that each lesson is observed, videotaped and analyzed. The research had three stages: the preparation phase (January – April 2007), the actual experimentation phase (May 2007), and the current retrospective analysis.

### 3.5.3 Participants

In the current study we follow in great detail one (out of three) sixth grade class (age 11–12, n=25) in a science-focused magnet primary school in Haifa. Most of the students come from affluent background and participated in the *Connections* EDA lessons in fourth and fifth grades. These previous encounters made them fluent with the software and basic informal statistical ideas, language, skills and perspectives.

### 3.5.4 Analysis

To assess students' learning we use video recordings of all sessions, researcher's observations, interviews of selected students and teachers, and students' artifacts (notebooks and project reports). We also administered an IIR pre- and post-tests. These tests included 10 items: Two were taken from the released items of 1995 TIMSS for eight-grade and eight items were adapted and modified to sixth grade from the ARTIST bank (https://app.gen.umn.edu/artist) and the interview protocol of Zieffler, Garfield, delMas, & Gould (2007).

The analysis of the videotapes is based on interpretive microanalysis (see, for example, Meira, 1998): A qualitative detailed analysis of the protocols, taking into account verbal, gestural and symbolic actions within the situations in which they occurred. The goal of such an analysis is to infer and trace the development of cognitive structures and the sociocultural processes of understanding and learning. Two stages are used to validate the analysis, one within the researchers' team and one with additional researchers in education, who have no involvement with the data (triangulation in the sense of Schoenfeld, 1994). In both stages the researchers discuss, present, and advance and/or reject hypotheses, interpretations, and inferences about the students' cognitive structures. Advancing or rejecting an interpretation requires: (a) providing as many pieces of evidence as possible (including past and/or future episodes and all sources of data as described earlier) and (b) attempting to produce equally strong alternative interpretations based on the available evidence. The final report will include cases in which the two analyses are not in full agreement, and points of doubt or rejection are not refuted or resolved by iterative analysis of the data.

## 4. Results
## Cases to be Discussed in SRTL-5

During the SRTL-5 presentation we shall discuss four brief video segments that were chosen to present typical ways of sixth graders' IIR. The cases are interspersed along the IIR instructional sequence and focus mainly on Oded and his team: Gal, or Asaf and Elad. The first video comes from a guided activity about random sampling, followed by two data

investigations based on repeated random samples. The fourth video vignette is part of the team's final presentation of their “research project”.

### 4.1 First Case: Random Sampling Guided Activity

### Oded and Gal

This is one of the first activities in the IIR learning trajectory. To illustrate the need for random selection we used the “Stringing Student Along” activity (Shaughnessy & Chance, 2005, pp. 43-44). A bag of 25 different length strings between 2 to 35 centimeters is used for drawing ten strings with replacement. Students in couples draw 2-3 samples of size 10 and compute the average of the length in the samples. The goal is to estimate the mean string length in the entire population.

Averages of the randomly drawn samples of strings are drawn on a dot plot by each team as well as collected on the classroom board from all teams. Having discussed the estimated parameter, students take all the strings out of the bag, measure their lengths, and compute the actual average length for the population of strings. They are surprised to find out that actual parameter is much smaller than their predicted parameter since this sampling method is biased. Longer strings are more likely to be chosen. As a result, this procedure over-represents the longer strings. The students are then asked to suggest different, better sampling methods. One method suggested by students and used in our follow-up activity is to use instead of actual strings, a set of 25 equal size small paper notes that have the strings length written on them. These notes are folded and drawn randomly from a hat.

The following interaction took place after Oded and Gal drew two random samples of ten paper notes each. They calculated the averages and summarized their data from previous lesson (biased sampling) and current lesson (random sampling) on a dot plot (Figure 6).

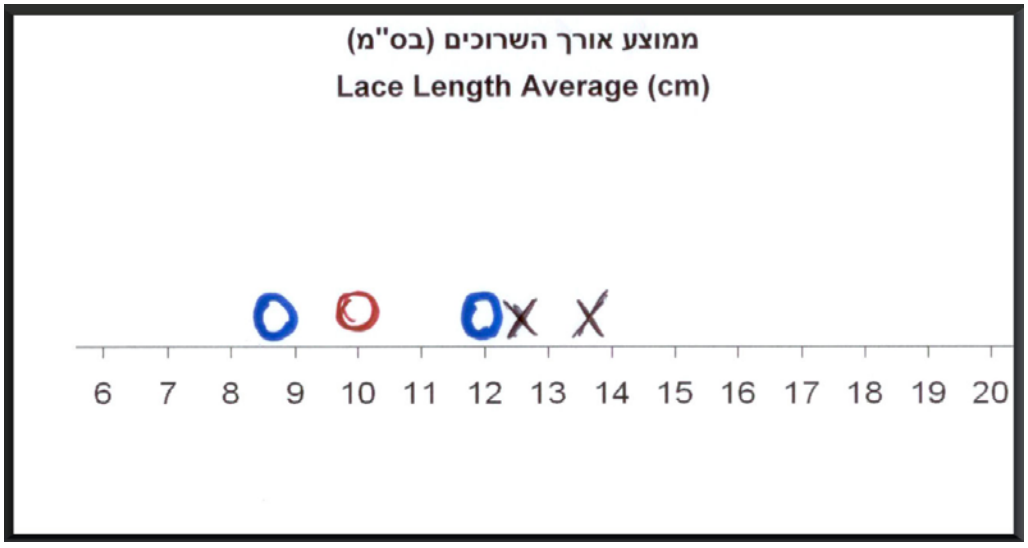


*Figure 6*: A dot plot drawn by Gal and Oded of random sample averages (n=10, blue circles), biased sample averages (n=10, black X-s), and a population average length of 25 strings (red circle).

| | | |
|---|---|---|
| 1 | Gal: | [Reading question 6.] How close were the notes sample averages to the population average compared to the laces sample averages? Explain. |
| 2 | Gal: | The [paper notes sample] averages are not so close to the laces [sample] averages [Figure 6], since there was a tendency to take the long laces. |
| 3 | Einat: | And in this [notes] sample? |
| 4 | Gal: | It was very random because there were no bigger or smaller notes… Since there was a "problem" that there were, for example, three 2's and one 34 [centimeters string]. |
| 5 | Oded: | Since the notes weren't in different sizes, we had no tendency to take the bigger ones, not like the laces. It is easier to pick the long laces, which probably made the results larger. |
| 6 | Gal: | And less close to the real average [Figure 6]. |
| 7 | Oded: | [Reading question 8.] What are the differences between the two sampling's methods? Is one method preferable than the other to find the real average of the lace population with greater confidence? |
| 8 | Oded: | In the lace sampling we had the problem of tendency, and in the notes sampling, the notes were in equal sizes, which made it more random … Therefore, I think that the notes sample is preferable, since the results were closer to the real average. |
| 9 | Gal: | The results are more reliable. |
| 10 | Einat: | What does reliable mean, in your opinion? |
| 11 | Oded: | It is something that you can trust, that you know more that … |
| 12 | Gal: | It's exactly correct. |
| 13 | Oded: | Yes, it's a correct result. It doesn't have to be related to results… If I have a reliable friend, it means that I have a friend that I can trust. |
| 14 | Gal: | [Reading question 9.] How could you implement the unbiased, random sampling in the lunch bags study [their first sampling activity]? Describe how you will randomly sample students to create a representative sample of the whole school population. |
| 15 | Oded: | [Hesitating] OK… It's impossible to… |
| 16 | Gal: | We can do it with a hat. Take all the class, cut [notes] exactly [in the same size], and then it will also be random. Well … [Turns to Einat:] Will it be random in the end? |
| 17 | Einat: | Tell me what random means. |
| 18 | Oded: | Yes, it's a good idea, because we can take a hat or a bag, take different small notes that are exactly the same size, and then the teacher will help us write the students' names. According to the sampling that I did [in the lunch bags study], we chose 10 kids from each grade. After we write all the kids in the class, we shall pick 10 notes… And whatever comes out – simply comes out. |

> We didn't have significant differences between boys and girls in the sample that I did with my friends [about the lunch bags]. We took equal number of boys and girls because the girls really wanted and also the boys [smiling]. But I don't think the gender has an effect…

An interview followed this activity in which Oded and Gal discussed a sampling distribution of the whole class (drawn ion the classroom board, see Figure 7). In this interview they demonstrated their understanding of various considerations related to sampling, randomness, bias and representativeness.

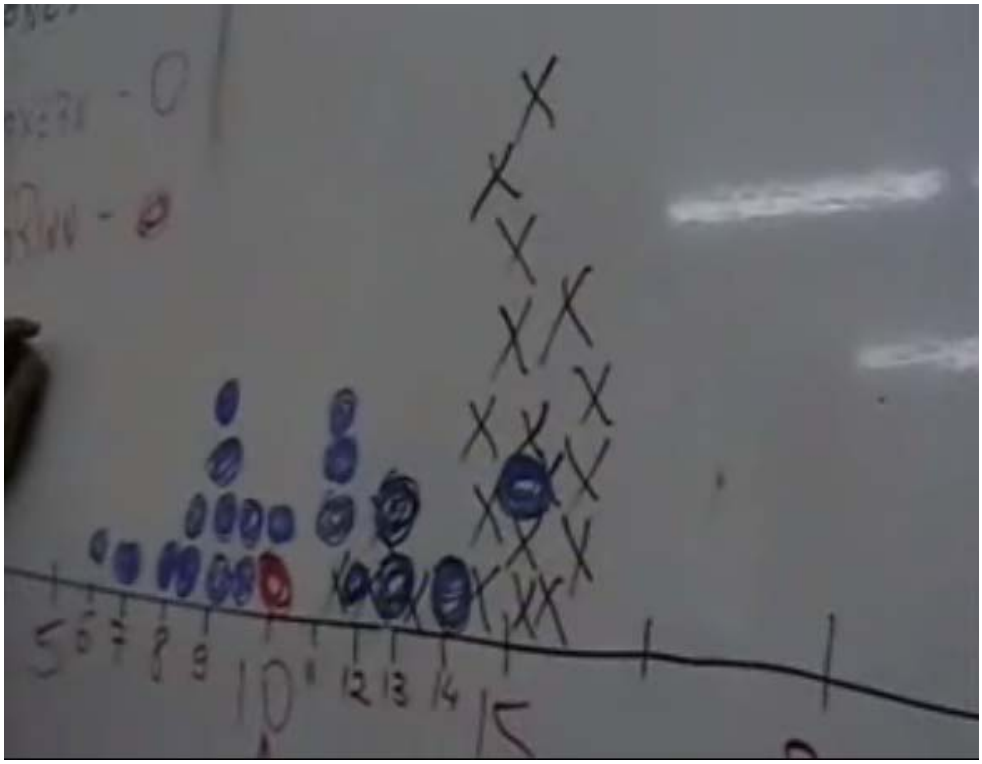

*Figure 7*: Sampling distributions of string sample average length (black X-s) and paper note sample average (blue circles). The red circle represents the population of 25 strings parameter.

### 5.2 Case 2: Homework Load – First Investigation
### Oded, Elad and Asaf

In order to collect real and relevant data, a 17-item questionnaire about gender and age, issues related to transfer from primary school to middle school (e.g., homework load), and sportsmanship (e.g., long jump results, favorite sport) was used. The data were collected from all students in grades 6 (n= 88) and 7 (n=118) in school to form the population to be investigated by students. The population data were fully entered to a *TinkerPlots* file that was never exposed to students. Students were only allowed to randomly sample from this file in order to infer the population parameter.

In this video vignette we observe Oded, Elad and Asaf in their first investigation of home work load of sixth and seventh grade. They randomly drew two samples of size 20 (10 from each grade level) using paper notes in a hat (the first sample is shown in Figure 8 while the second sample is presented in Figures 9 and 10). The following discussion took place when they try to interpret the results and infer about the population characteristics.

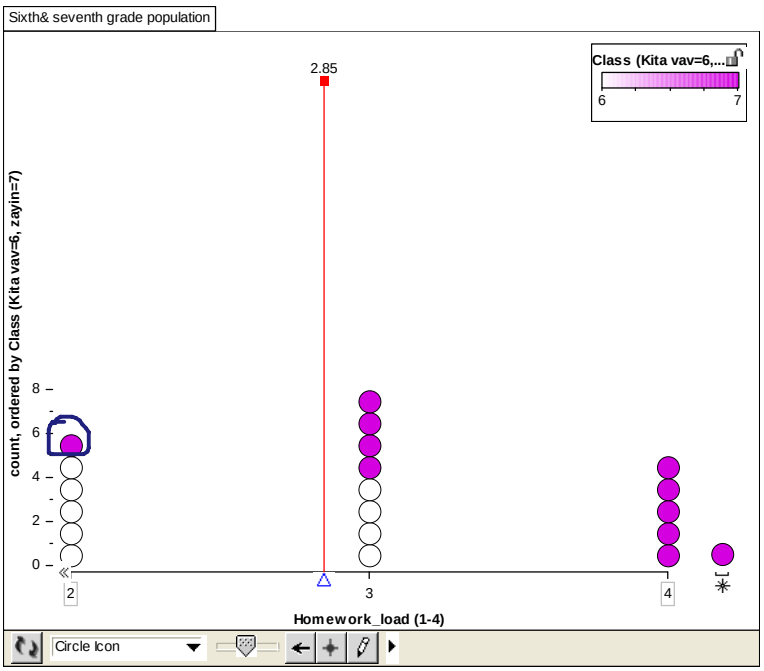


*Figure 8*: First sample (n=20) in Elad, Asaf and Oded's first investigation of homework load in sixth and seventh grades.

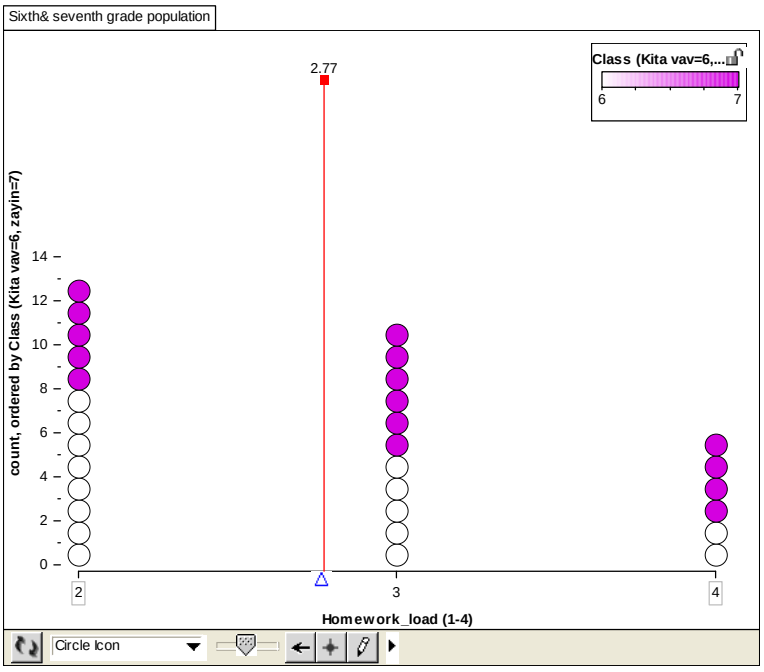


*Figure 9*: Second sample (n=30) in Elad, Asaf and Oded's first investigation of homework load.

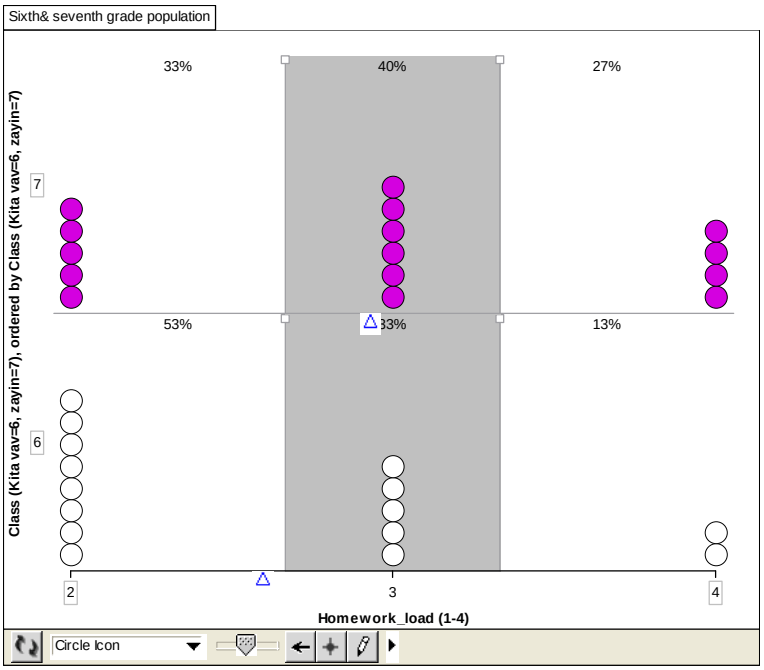


*Figure 10*: Second sample (n=30) with dividers
in Elad, Asaf and Oded's first investigation of homework load.

1 Elad: We are investigating if there is an association between grade and sense of homework load.

2 Oded: [Interpreting the first sample graph, Figure 8.]… Most of sixth graders chose between 2 and 3, which is a smaller homework load. … and seventh graders are mostly between 3 and 4. Therefore, we concluded that seventh graders have bigger homework load, except for one outlier from seventh grade that we marked, who is either a good student or doesn't do homework…

3 Oded: [Interpreting the second sample graph, Figure 9.] Our conclusion here is relatively similar… Although more sixth graders have high load, they are more or less in the same situation relatively to seventh graders. Perhaps, we have more weak students in this sample, which change the data…

4 Asaf [Interpreting the second sample plot with dividers, Figure 10.] We can actually see that seventh graders feel that the homework load is bigger than sixth graders. Most of… 53% of the sixth graders that we sampled think that their homework are quite easy while 40% of the seventh graders … that is the majority … actually think that the homework are harder… We can actually see that the homework load in seventh grade is different.

5 Einat: How confident are you in the inferences made from this sample? Do you think it is representative? Is this what really happens, or not necessarily?

6 Asaf: I think it's representative… It's also reasonable to assume that the load in seventh grade is larger and heavier than in sixth grade… [This conclusion is] not related to the sample.

| | | |
|---|---|---|
| 7 | Einat: | Aha. OK, so your inference is based on… The sample and also on how reasonable the conclusion is. |
| 8 | Asaf: | Yes |
| 9 | Einat: | If you got a result that is not reasonable in your opinion, what would you think? |
| 10 | Asaf: | I would think that the sample is not reliable or… |
| 11 | Einat: | And then, what would you do? |
| 12 | Elad: | Wait a minute, but maybe we didn't choose the sample in the most random way… We should "interrogate" ourselves to know if we really chose randomly. |
| 13 | Einat: | And what if you find that it was still random? |
| 14 | Elad: | Maybe we could take a look at the census and what's in it… |
| 15 | Einat: | But let's say that you can't look at the census… |
| 16 | Elad: | Yes… this [looking at the census] is really the shortest way, so I would enlarge the sample a little bit more. |
| 17 | Einat: | By how many, for instance? |
| 18 | Elad: | Twenty from each grade. |
| 19 | Einat: | Do you think that there is another way to strengthen the conclusions? |
| 20 | Elad: | I don't know… |
| 21 | Elad: | OK. Oded, maybe you have an idea? |
| 22 | Oded: | We can combine this sample with the previous sample, actually … mix the samples. And if we combine them, our conclusion will likely be more reliable. Since there may be children that are worse or better students, and the worse students may have much more or much less homework load …. If we combine the two samples it will likely expand our knowledge about sixth and seventh graders. However, it's possible that we get the same children in the combined sample. But nevertheless I think that we also have to enlarge the sample… |
| 23 | Einat: | Enlarge? |
| 24 | Oded: | Yes…. Well, I don't really know… |

### 5.3 Case 3: Long Jump – Second Investigation
### Oded, Elad and Asaf

Unsatisfied from the triviality of their first investigation, the three students decide to look for a more challenging and interesting topic: long jump results. This time they draw two samples of size 20 from the population using *TinkerPlots*' random generator (the first sample is presented in Figures 11 and 12, the second sample in Figure 13). The following discussion took place during this investigation.

| | | |
|---|---|---|
| 1 | Asaf: | Shall we choose to deal with a new research question? |
| 2 | Elad: | Yes. |
| 3 | Asaf: | Oded, Is it ok with you? |
| 4 | Oded: | Yes, the previous question was not a real question and our conclusion was quite self-evident. |
| 5 | Elad: | Shall we study something that is related to sixth and seventh grades and also to sports? |
| 6 | Asaf: | Yes, there will be more significant results. |
| 7 | Asaf: | Maybe, jogging? |
| 8 | Elad: | 60 [meter run]? |
| 9 | Asaf: | How about long jump? |
| 10 | Elad: | OK. And shall we also add “favorite sport”? |
| 11 | Oded: | [cynically] What shall we see? That most of the seventh graders that like for instance football jump like sixth graders that like basketball? |
| 12 | Elad: | [Looking at the second sample graph, Figure 11.] What do we study from this graph? |
| 13 | Asaf: | Hold on, hold on, that's interesting. The sixth grade results are relatively lower. |
| 14 | Elad: | Why? |
| 15 | Asaf: | They are more centralized in… |
| 16 | Elad: | They are higher on average... Let's add the averages [Average marks are added to the graph, Figure 11]. |
| 17 | Oded: | Yes, our [sixth grade] average is bigger than the seventh grade average. |
| 18 | Asaf: | How come? It doesn't make sense… |
| 19 | Asaf: | Perhaps we picked such and such girls… |
| 20 | Asaf: | Let's check the gender, the number of boys and girls. |
| 21 | Elad: | Oh... Good idea [Change the graph to represent gender, Figure 12.]. |
| 22 | Oded: | OK, you see that… |
| 23 | Elad: | Oy Vey, That's terrible… |
| 24 | Asaf: | Here [in sixth grade, Figure 12] we have one girl! |
| 25 | Asaf: | If we know that the boys jump farther than the girls, and we have more boys, it doesn't … |
| 26 | Elad: | Oy Vey… |
| 27 | Asaf: | Well, but we have chosen the sample randomly. |
| 28 | Elad: | That's right, and besides that, there is a majority of boys in the seventh grade… |

| | | |
|---|---|---|
| 29 | Elad: | There is only one way to see if it's correct, and do you know what it is? |
| 30 | Oded: | What? |
| 31 | Elad: | Enlarge the sample… Tadadam! |
| 32 | Oded: | Tadadam… |
| 33 | Asaf: | I agree that it can help. |
| 34 | Elad: | Yes. |
| 35 | Asaf: | How many shall we add? Ten more? |
| 36 | Einat: | What do you think about the idea of drawing a new sample? |
| 37 | Oded: | Let's draw a new sample and compare between the two samples. |
| 38 | Elad: | Do you mean a new sample with the same research question? |
| 39 | Einat: | Let me ask you something: Why do you think your conclusion is unreasonable? |
| 40 | Asaf: | Because the sixth grade long jump average can't be greater than the seventh grade average, even though we can explain that: boys probably jump farther than girls, and there is only one girl in the sixth grade sample. |
| 41 | Elad: | So maybe… shall we draw the same sample, but try to have an almost equal number of boys and girls? |
| 42 | Asaf: | How can you do that? |
| 43 | Elad: | I don't know… it might be like that in the next sample… but if we choose it one by one – the sample will be biased. |
| 44 | Asaf: | So, there are also disadvantages in a random sample. |

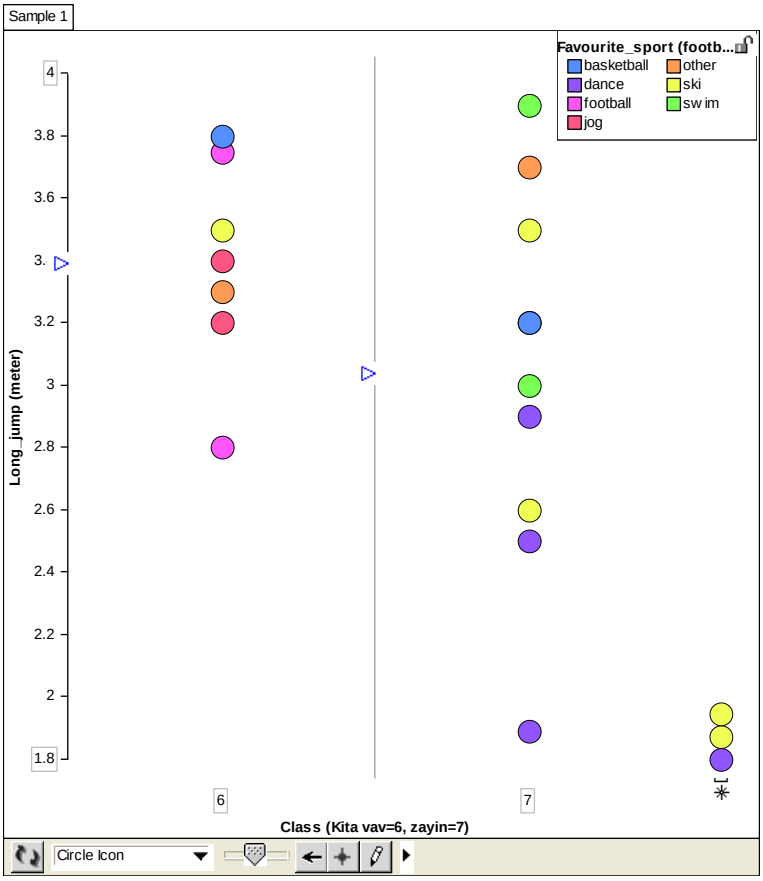


*Figure 11*: First sample (n=20) presented by favorite sport in Elad, Asaf and Oded's second investigation of long jump.

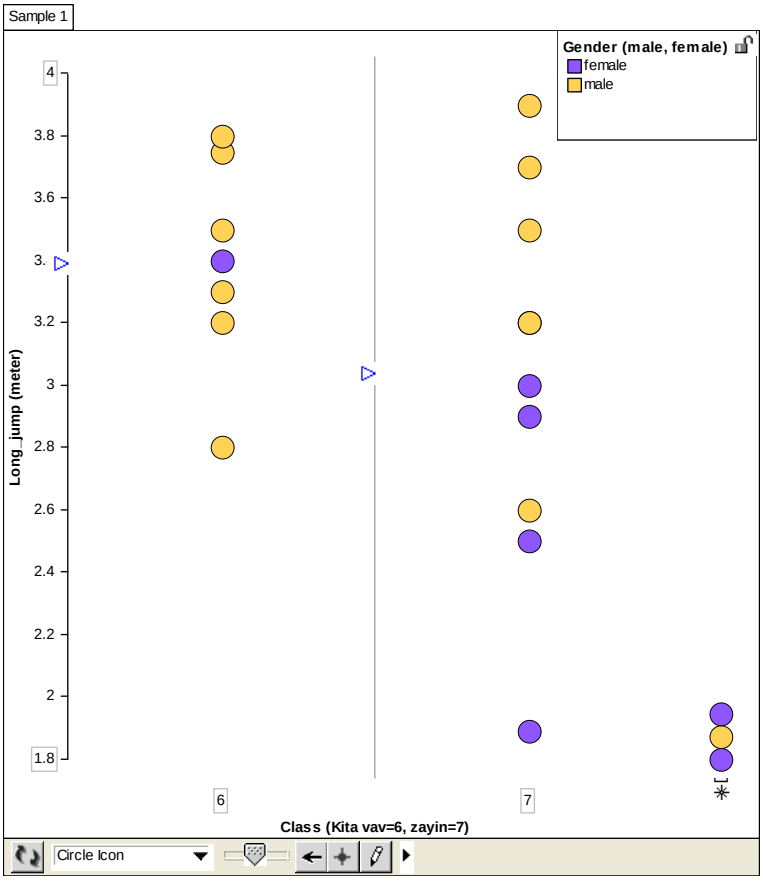


*Figure 12*: First sample (n=20) presented by gender
in Elad, Asaf and Oded's second investigation of long jump.

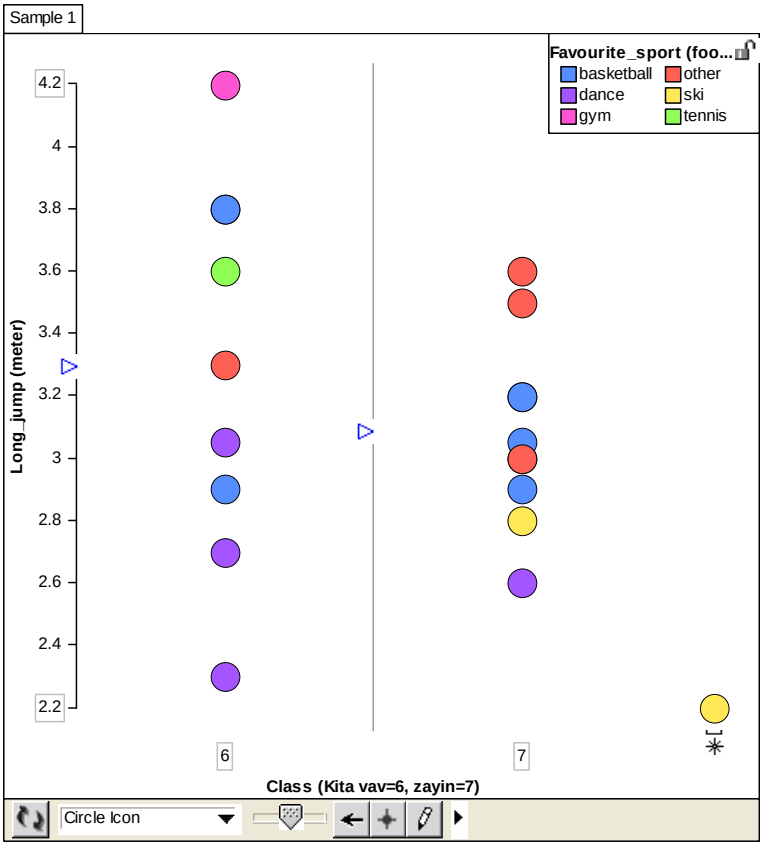


*Figure 13*: Second sample (n=20) in Elad, Asaf and Oded's second investigation of long jump.

### 5.4 Case 4: Long Jump – Final Presentation
### Oded, Elad and Asaf

Oded, Elad and Asaf continued struggling with the unexpected results of their second investigation. Two random samples that they drew made them infer that sixth graders jump better on average than seventh graders. They explain part of their reasoning in the final presentation of their research project.

| | | |
|---|---|---|
| 1 | Elad: | Our research questions are: What are the long jump results in grades 6 and 7? |
| 2 | Asaf: | And does the favorite sport affect these results? |
| 3 | Elad: | Is there an association between favorite sport and long jump results? |
| 4 | Oded: | Our hypothesis was that seventh-graders jump farther because they are apparently stronger and bigger. |
| 5 | Asaf: | And they are also more experienced than us. Therefore, we thought they'll jump farther. |
| 6 | Oded: | And we also hypothesized that favorite sports that include jumps, like basketball and gymnastics, will have a greater effect on long jump results. |
| 7 | Elad: | Well, here is the first sample [Figure 11]. |

| | | |
|---|---|---|
| 8 | Elad: | This is sixth-grade and this is seventh-grade. In this sample we see that the long jump average of sixth-grade is greater than the average of the seventh-grade. |
| 9 | Asaf: | It was very surprising! |
| 10 | Elad: | Here is the sixth-grade average and here is the seventh-grade average. |
| 11 | Asaf: | We were surprised to find out that the sixth-grade average was higher than seventh-grade … So we can actually say that our hypothesis "collapsed". |
| 12 | Oded: | This can simply happen because sixth-graders are nowadays more interested in sports than seventh-graders. |
| 13 | Asaf: | Another possible explanation can be found in another graph of this sample [Figure 12], which shows that in sixth-grade there are more boys than girls. The fact that boys probably jump farther than girls can explain that. |
| 14 | Elad: | And here is a second sample [Figure 13], which is in fact quite similar to the first sample since the sixth-grade average is also bigger than the seventh-grade average. |
| 15 | Asaf: | The two samples are in fact very similar and strengthen our conclusion. |
| 16 | Oded: | Well, in light of the similarity between the two samples, we found out that sixth-graders jumped farther than seventh-graders. We also saw that basketball and gymnastics really affected the long jump results. |
| 17 | Asaf: | From these two samples, we infer that the physical fitness in sixth-grade is probably better than in seventh-grade, or that more sixth-graders are engaged in sports that support long jump. We are certain about our inferences since the two samples were almost the same. Our level of confidence in our inference is at the level of 9 of 10. |
| 18 | Dani: | If you want to increase your confidence level in these conclusions, what will you do? |
| 19 | Asaf: | Get more samples… |
| 20 | Dani: | What else? |
| 21 | Asaf: | … or increase the sample size, or do something a little more complicated – a census of the whole population. |
| 22 | Asaf: | [The population graph, Figure 14, is exposed for the first time.] In fact, as we found, the seventh-grade average is 2.90 and the sixth-grade average is 3.07 [in the population]. We see a smaller change [difference], but still a change… One explanation can be the number of boys compared to the number of girls… |
| 23 | Dani: | Is your previous conclusion strengthens or weakens by what you see in the population? And Why? |
| 24 | Elad: | In fact, it strengthens it, but it also weakens it. |
| 25 | Dani: | How come? |
| 26 | Elad: | It strengthens it since we see that the sixth-grade average is really bigger |

than the seven-grade average, but on the other hand it weakens it since the gap between the two averages here is not so big.

27 Dani: What do you think about this whole process of sampling and inference that you went through?

28 Oded: I think that perhaps the [first] sample [Figure 12] was biased since we got different numbers of boys and girls …

29 Dani: What will you do next time when you infer from a sample?

30 Asaf: Maybe we can ask equal number of boys and girls… Otherwise the conclusion is not reliable.

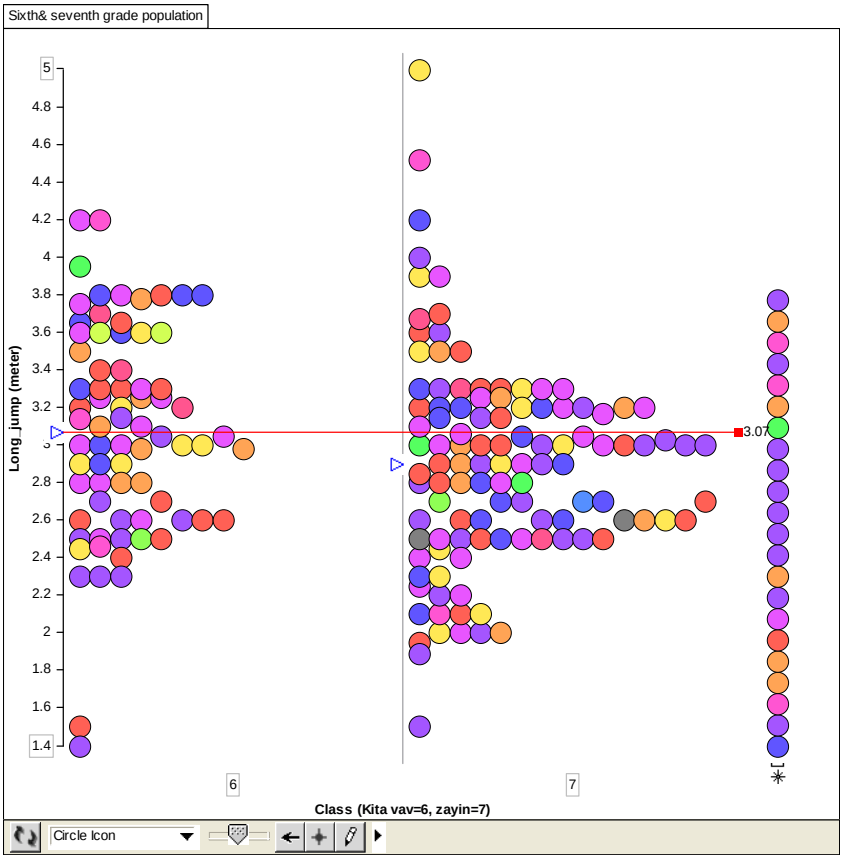


*Figure 14*: A population long jump graph prepared by Elad, Asaf and Oded during their final presentation.

## 6. Discussion

The following issues will be included in the discussion:

- The critical importance of longitudinal studies to understand students' emergent IIR and its development
- The importance of classroom-based (natural) research to understand the complex ways in which IIR is emerging in the sociocultural setting of the class
- The reciprocal relationships between the continuously changing IIR framework and the design and analysis

- The role of language about uncertainty (e.g., random, biased, confidence level) in scaffolding students' emerging IIR.
- Surprising and complex problem situations (e.g., long jump) in which data contradicts model and prior believes are useful
- The complex role of context: data-based claims are often mixed with contextual-based claims (Challenge: study context role in IIR)
- The current study challenges some of the "classical" results of the stat ed research literature, for instance:
    - Students using averages to compare groups
    - Students using a combination of local, semi-global and global methods to interpret graphs
- Complexity and ambiguities are part of IIR learning and therefore have to be addressed and not avoided (Challenge: how to do that?)
- Suggested sequence (concepts and activities) for a research-based IIR learning trajectory (Gil 2007)
- Such sequences (however) have to be re-examined before they go to class (curricula, standards, …)
- Shoulder to shoulder teachers' work with researchers is one way of making IIR class happen.
- Our understandings as researchers evolved to deeply appreciate sampling and inference as integral part of the traditional EDA, starting at early age (Challenge: affordances? modeling?)
- Proposed extended IIR inquiry cycle (Gil, 2007)
- Samples are drawn from "tangible" population
- Continuing intellectually-motivated and positive statistical experiences of students during schooling years are important
- TinkerPlots as an essential student's reasoning tool (Challenge: Grades 2-8?)
- The importance of norms and routines to developing critical statistical reasoning

## 7. Limitations

- Even when a phenomenon seems important and the data interpretation was validated and agreed upon, the question of the idiosyncrasy of the identified phenomenon may remain open.
- Therefore, in the current study, the data and interpretations from students in other classes assist in checking for generalizability of the phenomena, as well as offered for competing interpretations by others.
- Experimental setting – What about "normal" classes and schools?

## 8. Summary

We are pleased to be able to present at the SRTL-5 Forum several segments of the 2007 data. A preliminary qualitative analysis of these rich data will be presented and discussed. The presentation will be mostly qualitative, demonstrating both the analysis and outcomes. We hope for an interactive discussion on four aspects of the study: a) the results in terms of what can be learned about primary students' negotiating meanings of IIR in the context of statistical problem solving processes aided by carefully designed learning trajectory and *TinkerPlots*; b) the various implications of the study; c) the improvement of the IIR theoretical framework suggested in this paper (Figure 1); and d) the suggestion of future research directions.

## 9. Acknowledgements

We gratefully acknowledge the valuable contributions of Shira Yuval Adler to the preparation of this paper and the video segments that are part of it.

## 10. References

Abelson, R. P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum.

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to Reason about Distributions. In D. Ben-Zvi and J. Garfield, *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*, pp. 147–168. Dordrecht, Netherlands: Kluwer Academic Publishers.

Bakker, A., Derry, J., & Konold, C. (2006). Technology to support diagrammatic reasoning about centre and variation. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil [CD-ROM]*. Voorburg, the Netherlands: International Statistical Institute.

Ben-Zvi, D. (2006). Use of TinkerPlots by Fourth Graders to Reason Statistically. In A. Rossman and B. Chance (Editors), *Proceedings of the Seventh International Conference on Teaching of Statistics* (CD-ROM), Salvador, Bahia, Brazil, 2-7 July, 2006. Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., Garfield, J. B., & Zieffler, A. (2006). Research in the Statistics Classroom: Learning from Teaching Experiments. In G. Burrill and P. C. Elliott (Eds.), *Thinking and Reasoning with Data and Chance: 68th NCTM Yearbook*, 467-482. Reston, Va.: National Council of Teachers of Mathematics (NCTM).

Ben-Moshe, O. (2007). *Developing fourth-grade students' statistical reasoning about distribution with TinkerPlots software* (in Hebrew). University of Haifa: Unpublished Thesis.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education 18*, 382-393.

Gil, E. (2007). *Design considerations of research-based learning environment intended to develop six graders' informal inferential reasoning.* SRTL-5 Poster.

Kirschner, P. A., Buckingham Shum, S. J., & Carr, C. S. (Eds.) (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making.* London: Springer-Verlag.

Meira, L. (1998). Making sense of instructional devices: The emergence of transparency in mathematical activity. *Journal for Research in Mathematics Education, 29*, 121–142.

Moore, D. (2000). *The Basic Practice of Statistics*. New York: W. H. Freeman and Company.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman and B. Chance (Editors), *Proceedings of the Seventh International Conference on Teaching of Statistics* (CD-ROM), Salvador, Bahia, Brazil, 2-7 July, 2006. Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., Hammerman, J., & Konold, C. (2006). *Exploring Informal Inference with Interactive Visualization Software*. Proceedings of ICOTS-7.

Schoenfeld, A. H. (1994). Some notes on the enterprise (research in collegiate mathematics education, that is). *Conference Board of the Mathematical Sciences Issues in Mathematics Education, 4*, 1–19.

Schwarz, B. B., Neuman, Y. & Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity: An empirical study. *The Journal of the Learning Sciences, 12*(2), 221-258.

Shaughnessy J. M., & Chance, B. (2005). *Statistical questions from the classroom*. NCTM.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (eds.), *International Handbook of Mathematics Education I*, 205-237. Dordrecht, Netherlands: Kluwer.

TIMSS (1995). http://timss.bc.edu/timss1995i/TIMSSPDF/AMitems.pdf.

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

Zieffler, A. S., Garfield, J., delMas, R., & Gould, R. (2007). *Studying the development of college students' informal reasoning about statistical inference*. SRTL-5.