



HAL
open science

Covariational reasoning patterns exhibited by high school students during the implementation of a hypothetical learning trajectory

Miguel Medina, Ernesto Sanchez, Eleazar Silvestre

► **To cite this version:**

Miguel Medina, Ernesto Sanchez, Eleazar Silvestre. Covariational reasoning patterns exhibited by high school students during the implementation of a hypothetical learning trajectory. Eleventh Congress of the European Society for Research in Mathematics Education (CERME11), Utrecht University, Feb 2019, Utrecht, Netherlands. hal-02412811

HAL Id: hal-02412811

<https://hal.science/hal-02412811>

Submitted on 15 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Covariational reasoning patterns exhibited by high school students during the implementation of a hypothetical learning trajectory

Miguel Medina, Ernesto Sánchez and [Eleazar Silvestre](#)

Departamento de Matemática Educativa, CINVESTAV-IPN, Mexico

ingcivi_6@yahoo.com.mx, esanchez@cinvestav.mx, eleazar.silvestre@gmail.com

We report some evidences of an ongoing research on covariational reasoning with high school students concerned with the design and application of a Hypothetical Learning Trajectory (HLT) on problems of correlation and linear regression, as well as the role played by technology (Fathom) employed in the learning process. We observed student's performance was slightly better when using the software in the instructional sequence, such as identifying and using the correlation strength and elaborating more refined arguments when comparing their own estimated model against the least squares line.

Keywords: Covariational reasoning, linear regression and correlation, high school, Fathom

Context and research question

Research by psychologists and statistical educators have produced extensive and robust knowledge about students' covariational reasoning; for example, Zieffler and Garfield (2009, p. 11) summarized the general research findings in six statements: students are often, 1) significantly influenced by their personal beliefs with respect to their covariational judgments; 2) prone to believe that correlation exists where there's actually none; 3) only considering data that support the joint presence of variables but ignore those referring to joint absence of variables; 4) greatly challenged to grasp negative correlation than dealing with a positive one; 5) disposed to make estimates below the actual correlation's strength; and 6), susceptible to imply causal relationships when dealing with covariation tasks. In contrast, it's more difficult to find research reports on how to produce progressive changes in students' covariational reasoning as a result of teaching interventions in classroom, including why such changes would occur. In a recent literature review performed by the authors during the past 10 years, only five reports were found that include teaching interventions, two of which were carried out with tertiary level students (Inzunza, 2016, McLaren, 2012), two at high school level (Gil & Gibbs, 2017; Dierdorff, Bakker, Eijkelhof & van Maanen, 2011) and one at elementary level (Fitzallen, 2012). Such amount of investigations that report on the effects of instruction in students' covariational reasoning reflect Shaughnessy and others' recommendations for future statistical educational research, who highlighted the need to carry out teaching experiments and document the effects of instruction. On the other hand, regarding research on the development of scientific thinking, students' covariational reasoning has been widely investigated at the elementary and middle school levels (Zimmerman, 2007) but rarely at the high school level (even considering replica with adult subjects); in addition, the attended statistical content in most of these studies is mainly restricted to the association between qualitative variables. Thus, the results of our research aim to strengthen the body of knowledge on covariational reasoning within this particular trend.

In Mexico, the topic of numerical bivariate data, correlation and line of better fit is prescribed for most high school curricula, in which is common to find the use of dynamic statistical technology tools (e.g., Excel or Fathom) as a resource that could play a fundamental role in producing progressive changes in students' reasoning; in particular, technology can help promote covariational reasoning because it allows students to optimize the time it takes to make a graph and relieves the burden of making complex calculations. Given this context, the purpose of this study is to identify patterns in students' reasoning emerged from the implementation of a developing hypothetical learning trajectory about linear regression and correlation; we propose a research question consistent with the following analysis strategy: to determine and compare reasoning patterns with some previously detected in an exploratory-diagnostic study where no teaching intervention was performed. Therefore, the new progressive patterns provide cues to the achievements gained during the new teaching strategy, so that our research question becomes: *What reasoning patterns emerge in students' solutions to problems of correlation and line of best fit during and as a result of student participation in a learning trajectory that includes extensive use of Fathom?*

Conceptual framework

We rely on three fundamental ideas: (1) *covariational reasoning*, (2) the notion of *aggregate*, and (3) the mechanism of a *hypothetical learning trajectory*. Covariational reasoning is characterized as thinking process carried out by the subjects related to judgments and interpretations about the relationship that exists between two variables. Carlson, Jacobs, Coe, Larsen & Hsu (2002) indicate that covariational reasoning is defined as the cognitive activities included in the coordination of two quantitative variables while attending to the ways in which they change one with respect to the other; this type of reasoning also entails understanding many features of variation in which variation with respect to the linear fit model is highlighted. Some studies discuss the modeling of variation and the importance of considering both explained and unexplained variation when data is explored; these two characteristics of variation are of particular interest in the study of the linear regression between two quantitative variables because the isolation and modeling of aspects of variation allows to carry out predictions, explanations and control, as well as to question why variation occurs resulting in the search for its causes.

In a linear regression and correlation context, the notion of aggregate is related to the form in which learners perceive a bi-variate data set through its representation in a scatterplot, and it's also related with the objectives of interpreting the possible relation that exists between the variables that generate the data, propose a linear model that best fits the data and estimate new values for the response variable within the available data set. Bakker (2004) and Dierdrop et al. (2011) highlighted the importance of such notion in students' learning process since it allows for drawing conclusions about regression and correlation given that it is through this reasoning that one can identify trends that extend beyond the correlated data; that is, to assume the notion of aggregate as to set a structure opposed to the attributes that could be addressed to aisled data points.

Finally, the notion of an HLT was introduced by Simon (1995) as an essential component of a constructivist teaching and learning posture. The mechanism involves the selection and dynamic interaction between three main components: learning objectives, a set of mathematical tasks, and the associated learning hypotheses as a prediction of students' actions and reasoning in response to

the activities. We intend to employ such mechanism in our research both as a device for planning didactic interventions as well as a research vehicle, which requires a cycling revisiting of the three mentioned elements grounded in students' practices and reasoning patterns.

Method

The previous study, constructing the HLT. Learning activities of the HLT are based on a questionnaire applied in 2016 to two groups of high school students enrolled in their senior year (17-18 years) in a public school; a total of 96 students participated in the study (Medina, Olay & Sánchez, 2016). From the analysis carried out in this exploratory test, students' responses to the problems (described up next) exhibited the following main learning obstacles:

- Students do not assume the data set as a whole, as an aggregate with its own properties, but as isolated cases with independent ones; to draw a line of best fit, they do not consider the variation that exists between the data and the position the model should have, and tend to only consider some points of the data available (such as the first and last ones, or the highest and lowest values of the set).
- The idea of creating a diagram to represent the bivariate data does not arise intuitively; students don't clearly identify properties such as the intensity of the relationship between the variables nor the meaning of the data pairs; and when facing prediction problems, do not use statistical concepts but rely heavily on arithmetic procedures.

Based on the difficulties encountered in this study, learning hypothesis for the design of the HLT were developed: students should be able to convert data collected in tables to a graphic representation (scatter plot diagram); reflect on how to use all data when proposing their line of best fit; understand and discuss the meaning of the line of best fit as the line that minimizes the estimation error; consider the development of a scatterplot as a necessary tool for the analysis of the relationship between two variables; make hypotheses about the relationship between two variables (graduated: none, little, lots of) based on the analysis of a scatter plot diagram; and obtain the least squares linear model to determine the value of the requested response variable. These hypotheses were condensed in two main learning objectives, (1) the *learning about distributions of bivariate data*: which intends for students to identify what elements, like the form and the tendency of a cloud of data points, are determining to establish the type of relation that exists between two quantitative variables, and (2), the *learning about the variation of the data with respect to the linear adjustment model*: which intends for students to identify that the linear adjustment model has an intrinsic random character rooted in the data itself.

Participants, instruments and execution. The current version of the HLT was applied to a group of 40 students (15 males, 25 females; 17-18 years old), arranged in 20 pairs, also enrolled in their senior year on a public high school in Mexico City; they previously received instruction on some essential descriptive statistics topics (statistical graphs, tendency and variability measures) and a mild technical-procedural introduction to linear regression and correlation. The application of the experiment consisted of 4 sessions of approximately 2 hours each: students familiarized with Fathom's basic commands with an instruction activity during the first session; in session 2, students faced the *gas consumption* problem [adapted from Moore, 1998; item (i), figure 1] in which they

proposed their line of best fit using the software, as well as analysis of some characteristics of the possible correlation between the variables (intensity and direction); such evidence which allowed the researchers to produce detailed descriptions of the visual analysis performed by the students and their justifications.

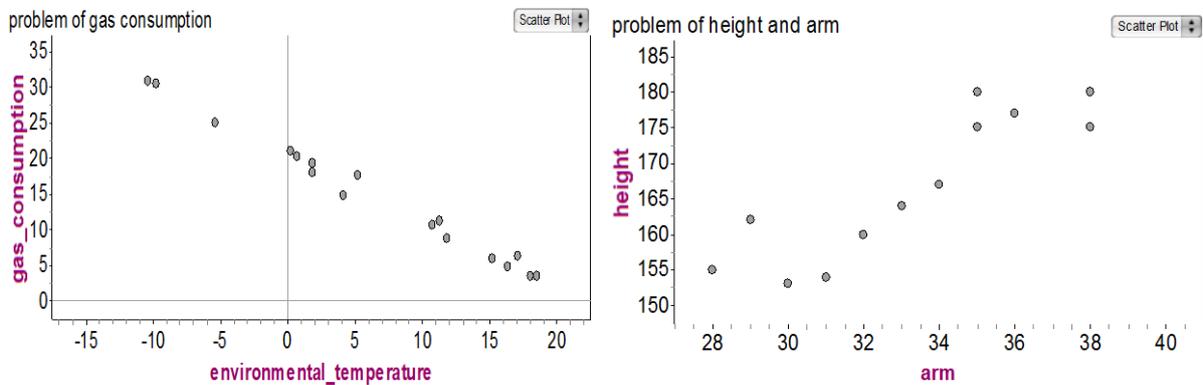


Figure1: Data used in the gas consumption problem [(i), left], and in the arm and height measures problem [(ii), right]

This is, students were asked about the form of the distribution of the data in the scatter plot, with which they could propose their own linear model that best fit the data (all of the instruction required the use of Fathom). In session 3, a further investigation of this problem was performed by adding a prediction item and a comparative analysis of the least squares line provided by the software against the students’ own model. Finally, in session 4, students faced the *height and arm measures* problem [item (ii), Figure 1], in which data was produced and taken by and from a sample of the participants. Once again, besides comparing students’ own model against the least squares line, the value corresponding to the sum of squares (errors) was provided through the software as a variability measure between the data and the linear models.

Data and analysis tools. Our primary source of data are students’ written responses to the activities, fieldnotes and video recordings. We rely on principles and basic coding techniques provided by the *Grounded Theory*, which emphasizes the creation of categories and emergent theory based on the data that is systematically and exhaustively collected during the investigation. Analysis of students’ responses consisted of identifying patterns exhibited in at least two pairs of students (open and axial coding); each one was given a code defined below, leading to ten emergent codes assigned into one of three groups: *inconsistent*, *favorable* or *neutral*; these adjectives refer to their relationship with normative covariational reasoning. Inconsistent codes correspond to reasoning features that we consider difficult the understanding of normative ideas embedded in covariational reasoning, while favorable codes prefigure or constitute partially aligned basis for normative reasoning considering that additional teaching support can help to overcome; neutral codes refer to those which we didn’t find reasons to be classified as neither of the former.

Results

Inconsistent codes are labelled A1-A5, favorable are labelled B1-B4, and the only neutral corresponds to C1; we now briefly describe and exemplify these codes using the notation “PX” to denote a specific pair of students’ response.

A1. *Illusion of linearity (or proportionality)*. When confronted with a prediction problem, it consists of proposing a plausible value of Y given a value x_0 , a specific point of the data set is chosen to formulate and apply a rule of three. An example of this reasoning comes with P8's response (gas consumption problem): "We applied a rule of three with the average temperature of 10.7 cause is the closest to 8, $\frac{10.7m^3}{10.7^\circ}(8^\circ) = 8m^3$ of gas".

A2. *Incidence in the maximum amount of points*. To determine the line of best fit, a line that goes through the maximum number of points is proposed, and the criterion is also used to judge the appropriateness of the model. This reasoning, exhibited in the gas consumption and temperature problem, is directly exemplified by P7: "...we positioned the line so it touched the maximum amount of points".

A3. *Attention to irregularity*. To describe a tendency or when judging the intensity of the correlation, points that deviate from a possible regularity or that are misaligned are noted and used as an argument. For example, when describing the tendency in the data presented in the scatterplot (height and arm measures problem), P4 responded "No [tendency is noted or present], there are only 4 points aligned".

A4. *Software references*. When describing a tendency in the data or justifying a proposal for the line of best fit, actions with the software are mentioned but no response to the actual question is given; as an example of such reasoning, P1 responded (gas consumption and temperature problem) "...we draw the line with the pointer until adjusting it".

A5. *Idiosyncratic*. Answers are given using terms that do not correspond to the common (statistical) usage; most likely, students assign a personal meaning that is not possible to deduce from their written statements. P18 exemplifies this rationale when describe their expectation about how gas consumption varies with the temperature by saying "it all should depend of the weather".

B1. *Describe the tendency*. Students formulate expressions that indicate if the tendency is either positive or negative. Such expressions can be widely general like "goes down/goes up", or much more specific "as long as X increments, Y diminishes". For example, when describing the type of tendency in both problems, P3 and P16 mentioned, respectively, "yes, the points are rising and forming a straight line", and "at a higher temperature the less [gas] consumption, and the lower the temperature the higher [gas] consumption".

B2. *Exclusive reference to the line*. The criterion consists of identifying the staggering of the points around a straight line but without describing its' direction; e.g., when describing the tendency in the data in the gas consumption problem, students like P20 provide answers like "yes, [the points group into] in a straight line".

B3. *Through the middle of the points*. This method consists of proposing a line that goes through the middle of the data set in the scatterplot; in some cases, it refers to getting an equal number of points on each side of the line (one half above and the other one under). As an example of such strategy, P5 justified their proposal of the best fit line in the height and arm measures problem by saying "we placed it in the middle of the points to adjust it the most".

B4. *Closeness*. The best fit line is the one considered as the closest to all points, in which “closeness” implies minimizing the sum of the residuals. This type of responses arises when students observe how the least square method is about using the software. An example of such thinking is provided by P1 in the height and arm measure problem: “...[we placed the line]so that all the points are the closest to it”, or by P9 when comparing and analyzing their own proposal of the best fit line: “...[in our case] the points are farer to the line, so the squares get bigger”.

C1. *Software use*. These responses only provide results or values taken from the software without adding comments or interpretations about them; for example, when P3 is trying to describe the intensity of the correlation in the height and arm measures problem, they just provide the value for r: “0.883844”.

Conclusions and discussion

We consider codes A1 and A2 as central when designing a teaching strategy. The term “illusion of linearity” has been adopted from van Doreen, de Book, Depaepe, Janssens & Verschaffel (2003) who rely on Freudenthal and define it as “the seduction to deal with each numerical relation as though it were linear”; the authors mention that such reasoning has been found in probability tasks as well as in arithmetic and algebraic domains. Casey (2014) pointed out that “many students required that their line [of best fit] go through the origin [...], they had a preconceived notion that all lines start there”, a similar reasoning we understand as a geometric form of the linearity illusion; however, it should be noted that both reasonings are not fully equivalent but mathematically related to each other. In our previous study (Medina et al., 2016), about 80% of the responses were coded as A1 while in the current one this percentage diminished to 40% (Table 1). We interpret that the often-inadequate exploitation of such resource goes beyond the overvaluation it receives in most elementary educational curricula (e.g., consider Darwin’s famous statement: “I have no faith in anything short of actual measurement and the rule of three”; such reasoning was quite common before Galton and others’ statistical contributions).

<i>Code</i>	<i>A1</i>	<i>B1</i>	<i>Others</i>
Gas consumption problem	8	3	9

Table 1: Incidence of codes in pairs of students given to the prediction item

The rationale behind code A2 (incidence in the maximum amount of points) was also found in Casey’s 2014 revision, in which she explains its presence as a consequence of previous learning gained in other mathematical domains such as analytic geometry. Besides addressing its possible causes, it is also convenient to discuss its possible implications. Students’ reasoning seems to partition the data into two parts, cases that rely directly into the linear model (explicit or imaginary) and those that doesn’t; that is, cases that belong or fulfill the model conditions and the ones that don’t. This lends support to the conjecture that at this point, students are still struggling to assume the data as an aggregate but at the most, as an aggregate partitioned in two.

The logic behind code A3 (attention to irregularity) could be expressed as a particular case of the former but focused exclusively on the points “outside” of the model. In doing so, the points that “deviate” from the model are considered as a different class with respect to the others, which may

be interpreted as a form to deal (or exclude) the notion of error involved in the use of statistical-inferential models, or even as expressions of overprivileged technical-deterministic reasoning (which could be addressed by promoting more reflection into the context of the data; e.g., by elucidating possible causes for the “lack of adjustment” of all the points).

As showed in Table 2, codes B1 (describe the tendency) and B2 (exclusive reference to the line) had an increased frequency when dealing with the second problem (final stage of the HLT), and a significant increase was also evident for code A3. The bigger frequency for code B1 might be attributed to students gained experience in this point of the trajectory despite the fact of dealing with a negative correlation; however, this feature might also have provoked that A3 appeared more often because of attending a more disperse data set.

<i>Codes</i>	<i>B1</i>	<i>B2</i>	<i>A3</i>	<i>Others</i>
Gas consumption problem	2	4	1	13
Height and arm problem	5	5	5	5

Table 2: Incidence of codes in pairs of students when describing the tendency in the data

As previously mentioned, codes B3 and B4 appeared as criteria of students when attempting to elaborate a linear model of their own (Table 3). Such codes were absent in our 2016 study, which led us to believe that their emergence might be motivated by the incorporation of the software to the instruction sequence. Specifically, adding a movable line and then comparing it against the least squares line seemed to have enable students identify how the former “goes through all of the points”; also, displaying and interpreting the sum of the squares through the software helped to capitalize the idea of the line of the best fit as the one that best approximates all points in the data set.

<i>Codes</i>	<i>B3</i>	<i>B4</i>	<i>Others</i>
Gas consumption problem	2	0	18
Height and arm problem	2	6	8

Table 3: Incidence of codes in pairs of students when elaborating a line of best fit on their own

To summarize and conclude this report, codes A1 and A2 are considered patterns that evidence students’ struggle when considering a data set as an aggregate in the context of linear regression and correlation problems. Codes B1 and B2 indicate a progress in students’ conceptions regarding their ability to conceive and identify a general structure subjacent to the data, although with great challenge when it comes to interpret and reconcile the random inferential nature of the model. The emergence of codes B3 and B4 suggest that the use of Fathom might have contributed to improve students’ understanding by helping them to associate the line of best fit to the one that minimizes the sum of errors or deviations between the points and the model. Finally, a major limitation of the HLT is the lack of incidence in promoting reflection upon the reasons of why such correlations might exist (or not); a new version of the HLT and future researches could target such deficiency by addressing students’ reasoning in aspects related to causality and variability.

References

- Bakker, A. (2004). Reasoning about Shape as a Pattern in Variability. *Statistics Education Research Journal*, 3(2), 64–83.
- Carlson, M., Jacobs, S., Coe, E., Larsen, S. & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events. *Journal for Research in Mathematics Education*, 33 (5), 352–378.
- Casey, S. (2014). Teachers' knowledge of students' conceptions and their development when learning linear regression. In K. Makar, B. de Sousa & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics. ICOTS9 Invited Paper-Refereed*. Voorburg, The Netherlands: International Statistical Institute.
- Dierdorff, A., Bakker, A., Eijkelhof, H.M.C. and van Maanen, J.A. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning*, 13, (pp. 132–151) (20 p.).
- van Doreen, W., de Boer, D., Depaepe, F., Janssens, D., Verschaffel, L. (2003). The illusion of linearity: Expanding the evidence towards probabilistic reasoning. *Educational Studies in Mathematics*, 53, 113–138.
- 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, pp. 290-297). Sydney, NSW: MERGA.
- Gil, E. and Gibbs, A. L. (2017). Promoting modelling and covariational reasoning among secondary school students in the context of Big Data. *Statistics Education Research Journal*, 16(2) pp. 163–189.
- Inzunza, S. (2016). Analysis of bivariate data in an environment based on applets and dynamic software. *Mathematics Education*, 28(3), 61–89.
- McLaren, C. (2012). Using the height and shoe size data to introduce correlation and regression. *Journal of Statistics Education* 20(3), 1–9.
- Medina, M. N., Olay, G. E., Sánchez, E. (2016). Razonamiento de estudiantes de bachillerato frente a un problema de regresión [Reasoning of high school students in a regression problem]. *Eutopia. Revista del Colegio de Ciencias y Humanidades para el Bachillerato*, 9(24), 71–78.
- Moore, DS (1998). *Basic Applied Statistics*. Antoni Bosch editor.
- Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26 (2), 114–145.
- Zieffler, A. S, & Garfield, J. (2009). Modeling the Growth of Students' Covariational Reasoning during an introductory statistics course. *Statistics Education Research Journal*, 8(1), 7–3.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review* 27, 172–223.