Visualisation of Lines of Best Fit

Michael Rudziewicz
Appalachian State University
michaelrudziewicz@gmail.com

Michael J. Bossé
Appalachian State University
bossemj@appstate.edu

Eric S. Marland
Appalachian State University
marlandes@appstate.edu

Gregory S. Rhoads
Appalachian State University
rhoadsgs@appstate.edu

Humans possess a remarkable ability to recognise both simple patterns such as shapes and handwriting and very complex patterns such as faces and landscapes. To investigate one small aspect of human pattern recognition, in this study participants position lines of "best fit" to two-dimensional scatter plots of data. The study investigates the variation in participants' fits and whether there is some consistent metric being used in fitting the lines. For example, is there a natural tendency toward fitting lines similarly to one of the standard regression lines: vertical, horizontal, or orthogonal. This study also investigates the effect of outliers on the line a participant fits to a scatter plot with a strong linear trend and provides guidance for future inquiries. Finally, this investigation considers some practical aspects for classroom teachers.

Keywords: Lines of Best Fit, Pattern Recognition, Cognition, Visualisation

Visualising Lines of Best Fit

Pattern recognition might be defined as the ability to process input, interpret information, and detect patterns about an environment. Many living organisms, including humans, have the ability to visually recognise patterns. Pattern recognition can be used to visualise trends in data. This paper seeks to explore how individuals visualise patterns (trends in data) by having them fit a line of best fit to multiple scatter plots. Those fits are then compared with each other and with standard regression lines (i.e., the vertical, horizontal, or orthogonal regression line) to determine if there are tendencies among these lines of fit. This investigation addresses a number of questions:

- What kinds of variation are in the lines people fit? That is, how consistent are participants in fitting lines?
- Do participants fit lines generally closer in proximity to one of the standard regression lines than the other two? Is it dependent on the specific data set?
- Would factors such as outliers or linear and non-linear data trends (e.g., linear, exponential, clustering, piecewise, and logarithmic) affect the selection of a line of best fit?

The results of this study provide insight of how individuals see patterns and construct lines of best fit. This paper also provides suggestions for further directions for research.

Recognising how students perceive lines of best fit may reveal to teachers that cognitive factors are at play. Students' lines of best fit are not merely good or bad; they perceive data in ways that may or may not align with mathematical convention. In conjunction with the findings of this study, practical issues affecting teaching are later posed.

Background Literature

Common Core Connections

The following Common Core mathematics content standards (National Governors Association for Best Practices & Council of Chief State School Officers, 2010) applicable to visualising lines of best fit include:

6.SP.B.5.C: Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.

6.SP.B.5.D: Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data was gathered.

8.SP.A.1: Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

8.SP.A.2: Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.

8.SP.A.3: Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.

HSS.ID.A.3: Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

HSS.ID.B.6: Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

HSS.ID.B.6.A: Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.

HSS.ID.B.6.B: Informally assess the fit of a function by plotting and analyzing residuals.

HSS.ID.B.6.C: Fit a linear function for a scatter plot that suggests a linear association.

HSS.ID.C.7: Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

HSS.ID.C.8: Compute (using technology) and interpret the correlation coefficient of a linear fit.

The notion of recognising the "context in which the data was gathered" is explicitly stated in respect to sixth grade standards regarding univariate data. However, in the eighth-grade standards regarding bivariate data, the important notion of the context from which the data was gathered is omitted; only an association within the data is considered. The eighth-grade standards mention "patterns of association between two quantities" and "relationships between two quantitative variables" and the recognition of linear and nonlinear associations. Unfortunately, this omission in the eighth-grade standards potentially decontextualises data and can possibly make the context of data seem irrelevant to students. Notably, recognising the context of the data is necessary for the justification of the appropriateness of fitting a line to a set of data. Fortunately, albeit sporadically, the notion of the context of the data returns in the high school standards. Numerous important concepts are addressed in the Common Core standards,

including: deviations from the overall pattern of the data (outliers); positive or negative association; linear and nonlinear associations; informally and formally fitting a straight line to the data; informally and formally assessing the model fit of linear and other appropriate functions to the data. Addressing these concepts in the classroom require further investigation and discussion; these follow later.

Visualising Data through Scatter Plots

Representations are constructs (e.g., artefacts, objects or devices, whether external or internal) through which individuals interpret and make sense of situations (Kaput, 1989), and, through which, relations are depicted, maintained, and acted upon (Olson & Campbell, 1993). Mathematical representations include, but are not restricted to, numeric (or tabular), symbolic (or algebraic), graphical, and verbal (oral or written) articulations of mathematical ideas.

Discussions of representations often distinguish between the notational schemata and characteristics associated with depicting a notion and the notion itself (Adu-Gyamfi, & Bossé, 2014; Adu-Gyamfi, Bossé, & Chandler, 2015; Adu-Gyamfi, Stiff, & Bossé, 2012; Bossé, Adu-Gyamfi, & Chandler, 2014). Various nomenclature is employed to denote this notational schemata including the format and operators of a representation (Ainsworth, 1999), the symbol scheme of a representation and its concretely realisable set of characters and rules for identifying and combining them (Kaput, 1987b), and the primitive elements, characters, or signs of a representation with their collection of permitted configurations (Goldin, 1987). While representations act as recognised symbol systems to communicate mathematical ideas or relationships (Duval, 2006; Goldin, 1987; Janvier, 1987; Kaput, 1987; Steinbring, 2006), in order for them to be usable, students must be able to decode, encode, and selectively combine the notational schemata in order to create a representational structure which correctly depicts a particular mathematical idea (Kaput, 1987; Lesh, Post, & Behr, 1987).

When students interact with representations, they can do so in two ways: syntactic elaboration denotes interacting with a representation by directly manipulating the notational schemata in the representations without reference to the meaning of the idea represented, and semantic elaboration denoting interacting with a representation based on the ideas represented, rather than the notational schemata themselves (Kaput, 1987). Syntactic elaboration often stunts a student's ability to interpret, interact with, and apply a representation.

Much is involved in students interpreting mathematical representations. Some speak of students interpreting a representation and, often unsatisfactorily, attempting to discern between valuable information and what can be ignored (e.g., Bossé, Adu-Gyamfi, & Chandler, 2014; Duval, 2006; Kaput, 1987b; Lesh, Post, & Behr, 1987; and Sternberg, 1984). Additionally, some representations present a greater number of confounding facts affecting both the interpretation of, and interaction with, the representation (Bossé, Adu-Gyamfi, & Chandler, 2014; Bossé, Adu-Gyamfi, & Cheetham, 2011).

In mathematics, visualisation is an important component in stimulating the mind, helping people see trends and patterns, and enabling them to "see the unseen," such as concepts, trends, and patterns. Arcavi (2003) provides the following definition for visualisation.

> "Visualization is the ability, the process and the product of creation, interpretation, use of and reflection upon pictures, images, diagrams, in our minds, on paper, or with technological tools, with the purpose of depicting and communicating information, thinking about and developing previously unknown ideas and advancing understandings." (p. 217)

Strong parallels clearly exist between Arcavi's definition of visualisation and previously mentioned characteristics of representations. Altogether, the act of visualising mathematical patterns leading to diagrammatic presentation can be seen as tantamount to developing a mathematical representation to depict a mathematical relationship.

While a scatter plot might look like a random assortment of points scattered on a graph, patterns often exist within them – or, at least, patterns can be induced by the student onto the data. For instance, a data set may be seen to possess a negative linear trend or a collection of points may describe the association between how long students studied for a test and the numeric grade each student received on the test. "The visual display of information enables students to 'see' the story, to envision some cause-effect relationships, and possibly to remember it vividly" (Arcavi 2003, pg. 218). By investigating participants' placement of lines of best fit, researchers may be able to determine how subjects visualise trends in different scatter plots and compare subjects' lines of best fit with commonly recognised regression lines (Mosteller, Siegel, Trapido, & Youtz, 1981.)

Linear Regression, Patterns, and Outliers

Through a scatter plot of data, a line of best fit can be used to represent that data. This is often accomplished by considering residuals. A residual, with respect to the commonly used vertical regression line, is the vertical distance between a datum point and the approximated line of best fit at the x-coordinate of that datum point (Embse, 1997). The goal of linear regression is to minimise the sum of the squares of the residuals, added as Euclidean distances (the square root of the sum of the squares). Vertical, horizontal, and orthogonal regressions minimise the sum of the squares of the vertical, horizontal, and orthogonal residuals respectively (Leng, Zhang, Kleinman, & Zhu, 2007).

While it is possible to perform a linear regression on any data set, this may not always be appropriate or helpful. Data sets are often found to have different patterns, such as linear trends, clusters, trends with outliers, as well as nonlinear trends. The standard linear regression may be seductive in its simplicity; it is fairly straight-forward to calculate, compared to other regressions, and is commonly used to interpret data (Motulsky & Ransas, 1987). However, if the data possesses a nonlinear trend or if specific information is known about measurement error in the independent variable, this linear regression may not be appropriate.

Outliers are individual values that fall outside an overall pattern (Moore, Notz, & Fligner, 2007). The addition of outliers can significantly change the regression line of a data set (Rousseeuw & Leroy, 1987). For instance, given a data set of five points with a strong linear relationship, a vertical regression line might demonstrate a very strong correlation. However, if one point is moved significantly away from its original position – becoming an outlier – the regression line may correspondingly change significantly and then, demonstrate a weak correlation.

Visualising Lines of Best Fit

Many decisions must be made by participants as they attempt to visualise a line of best fit through data. They must decide how appropriate a linear approximation may be to represent the trend of the data, and must also decide whether outliers should be considered in the development of the line of fit or should be omitted from consideration.

Mosteller, Siegel, Trapido, and Youtz (1981) investigated the proximity between a visualised line of best fit through scatter plots of sets of data and the associated vertical linear regression lines. Each participant was given a set of four scatter plots. They were also given a straight line printed on a transparency sheet. Participants were then asked to move the sheet with the line over the scatter plot until they were satisfied that the position of the line represented their visualised line of best fit. The slope and *y*-intercept of the visualised lines were estimated. The authors then compared the participants' average slope and intercept with the vertical regression line for each data set. This activity was replicated on three additional but different scatter plots. The results were that the averages of the participants' slopes and intercepts were close to the regression line for most of the data sets, with the exception of the data set with the weakest linear trend. The authors did not ask participants to visually fit a line to data sets that did not have a strong linear trend, to see how visualised lines would approximate or depart from regression lines in those cases.

The study presented in this paper expands on this previous work in a number of ways: it uses technology to find precise equations of participant generated visualised lines; it compares participants' visualised lines of best fit with vertical, horizontal, and orthogonal regression lines; and it makes use of multiple methods to make comparisons and analyse the results.

Methods

Expanding the methods of Mosteller, Siegel, Trapido, and Youtz (1981), a software program, The Geometer's Sketchpad (GSP) (www.keycurriculum.com), was employed to display scatter plots of data, allow participants to position a line through the scatter plots, and determine the function equations for these lines. With GSP's easy to use interface, the study was able to consider numerous data sets, some with stronger and weaker linear trends, some with clearly non-linear trends, and some with significant outliers.

Participants were 103 university students (male=69, female=34, with 95 such that $18 \leq age \leq 23$ and 8 such that $24 < age$) in various sections of a business calculus course in the south-eastern United States. From prerequisite courses (e.g., Math for Liberal Arts, College Algebra, or Precalculus), each participant had some knowledge and experience regarding calculating least squares regression lines using technology such as a graphing calculator or Excel. (Notably, calculating lines of best fit differs significantly from visualising lines of best fit through data plots.) All participation was voluntary, independent of course assessment or incentive, and performed in a computer lab overseen by the researchers.

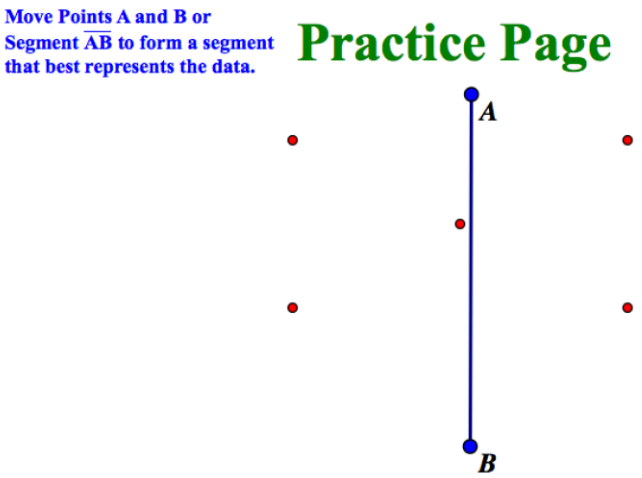

Figure 1. Practice Page

Since few participants in the class had experience with GSP, they were all provided instructions and practice regarding positioning a segment through a scatter plot of bivariate data to represent the line of best fit. (See Figure 1.) The actions required in GSP were simply to drag the endpoints of a line segment or the segment itself to best represent the trend they observed in each data set.

For each scatter plot, participants positioned their segment of best fit until they were satisfied with its position, then moved onto the next scatter plot. When lines were fitted to all scatter plots, the participant sent an email with their completed GSP file to a researcher. For each line fitted by participants, GSP was then used to find the equation of the line, with a four-decimal approximation of the slope and *y*-intercept.

Each scatter plot varied, containing linear patterns with different correlations, outliers and nonlinear trends. Some of the scatter plots were strongly linear with significantly large associated values for R^2 (Scatter Plots 1, 5, and 8). Some had one to three "intuitively obvious" outliers - albeit unmeasured as such, (Scatter Plots 2, 8, and 9). Thus, altogether, the set of scatter plots spanned: higher values for R^2 with no obvious outliers (Scatter Plots 1 and 5); higher values for R^2 with some "intuitively obvious" outliers (Scatter Plot 8); lower values for R^2 with no obvious outliers (Scatter Plots 3, 4, 6, and 7); and lower values for R^2 with some "intuitively obvious" outliers (Scatter Plots 2 and 9). These scatter plots are provided in the findings section later in this paper. It was anticipated that this set of scatter plots would sufficiently span possibilities to be investigated through the research questions.

The scatter plots on which participants were to position lines of best fit had axes hidden. This was done to ensure that the axes were not confounding facts to the participants; confounding facts could potentially lead to the distortion of the interpretation of the data alone.

Method 1

Each fitted line was compared to the calculated vertical, horizontal, and orthogonal regression lines to determine which regression line was closer to each fitted line. In order to determine and compare the proximity of the visualised line of best fit with these regression lines, the L-2 norm was calculated on the interval between the minimum data and maximum data value

in the set. If $f(x)$ is the fitted line by the participant and $g(x)$ is either the vertical, horizontal, or orthogonal regression line, we define the L-2 norm by $\left(\int_a^b (f(x) - g(x))^2 \, dx\right)^{1/2}$, where *a* is the minimum and *b* is the maximum value of the *x* coordinate of the data set.

Analysis for Method 1

For each scatter plot, a count of the number of participants whose L-2 norm was smallest for each of the three standard regression lines was determined. Then, the Chi-Square Test for Homogeneity was applied to determine if lines fitted by participants were closer to one regression line than the other regression lines. If p_1, p_2, p_3 are the probabilities that a randomly chosen participant will fit a line closer to the vertical, horizontal, or orthogonal respectively, the Chi-Square Test was setup as: H_o: $p_1=p_2=p_3$ and H_a: not all $p_i=1/3$. The null hypothesis, H_o, states that the lines fitted by participants are equally likely to be closer to any one of the three regression lines, and the alternative hypothesis, H_a, states that lines fitted by participants are not all equally preferred. The Chi-Square Statistic is represented by $\chi^2 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$ where the sum is taken over the three types of regression lines. The expected count is the expected number of participants fitting a line closer to that particular line, which is one-third of the total participants. The chi-squared test for homogeneity was used to find the p-value with 2 degrees of freedom. If $p < 0.05$, the standard level, the null hypothesis is rejected, which would mean for that particular scatterplot, a participant's line is not equally likely to be closer to the three regression lines. Otherwise, there is insufficient evidence to say the three lines are not equally likely.

Another statistical application applied in this study was a matched pair design that compared two subjects. Two regression lines were chosen and it was determined if a user's fitted line more closely approximated one regression line over the other for a given scatter plot. The following pairs were investigated: *H-V* = Horizontal – Vertical; *H-O* = Horizontal – Orthogonal; and *V-O* = Vertical – Orthogonal.

For each pair, the difference of the L-2 norms for the two lines was calculated. A Chi-Square Test was setup in a similar fashion as before, where, if *p* is the probability the difference of the L-2 norms is greater than 0, the null hypothesis is: H_o: $p=½$ and alternative hypothesis: H_a: $p\neq1/2$. The null hypothesis, states that the lines fitted by participants do not have a preference between the two regression lines in the pair, and the alternative hypothesis says that there is a preference between the two lines. A 95% confidence interval of the mean difference was also calculated for each matched pair for each scatterplot to give another way to determine if lines fitted by participants more closely approximated one regression line over another. Each confidence interval contained a lower and upper bound of the mean for the matched pair. Since it was reasonable to expect that each matched pair would not have a normal distribution, bootstrapping was used to find an average 95% confidence interval for the mean for each matched pair. Bootstrapping involves resampling with replacement from the original sample. One thousand new sample sets were created for each scatter plot, with one hundred and three samples in each set and a confidence interval was calculated for each set. An average confidence interval was then calculated for each matched pair, and the location of 0 with respect to this interval could provide evidence of one regression line in the pair being closer to the fitted lines than the other.

Method 2

Using slope intercept form to compare the participants' lines of fit with the regression lines may have some limitations. For instance, although the slope remains a solid comparative measure of the lines, a shift in the slope of a line significantly affected the *y*-intercept and even a slight shift in the slope was magnified greatly in the *y*-intercept. This is especially true when the data does not span the *y*-axis. In our case, all of the data was contained in the first quadrant. Thus, the dependency of the *y*-intercept on the slope of the line could skew the comparison of the participants' lines of fit with the regression lines.

Therefore, noting that the standard regression lines all passed through the centroid of the data, the orthogonal distance from the centroid to the data to the fitted line was considered a more stable measure for comparison. In addition to the distance, a sign was added to indicate whether the line was above or below the centroid. Analysing with respect to slope and the orthogonal distance from the participant's line to the centroid allowed for a set of observations and findings that are complementary to the findings based on the slope-intercept forms of the lines. Both sets of findings are shown in tandem through the discussion of the results in this investigation.

Results

The lines fit by the participants extended over a relatively broad range. Table 1 shows the slopes intercepts, and centroid distances of the lines. It is clear that the spread of the slopes is narrower than that of the intercepts which might be expected, since the intercept is a vertical point evaluated at the extreme edge of the range used and changes in the *y*-intercept are magnified by even slight changes in the slope of the respective line. In contrast, the distance from the fit lines and the centroid of the data is quite small, perhaps giving a more meaningful measure of the spread of the vertical placement than the intercept.

Table 1
Boxplot results from the slope, intercept, and distance from the Centroid of participants' lines.

Scatter Plot	Slope			Intercept			Distance from Centroid		
	Q1	Median	Q3	Q1	Median	Q3	Q1	Median	Q3
1	1.03	1.10	1.15	-0.95	-0.54	-0.14	-0.82	-0.03	-0.00
2	0.57	0.67	0.77	0.46	1.45	2.43	-1.71	-1.53	-1.09
3	1.06	1.28	1.42	-0.22	0.20	0.94	-0.80	-0.28	0.02
4	-1.28	-1.20	-1.00	21.01	21.35	22.24	-0.30	0.10	0.56
5	0.78	0.85	0.92	0.14	0.62	1.81	-0.93	-0.40	-0.02
6	0.72	0.81	0.95	0.50	0.52	1.45	-0.44	-0.13	0.23
7	0.89	0.99	1.06	-0.81	-0.31	-0.01	-0.11	0.37	0.20
8	-1.60	-1.40	-1.22	8.28	9.20	10.00	-0.20	-0.04	0.29
9	1.04	1.18	1.30	-1.24	-0.60	0.04	-0.16	-0.05	0.12

Employing research Method 1, Table 2 provides the number of the fitted lines that are closer to each particular vertical, horizontal, or orthogonal regression line for each scatter plot along with the total count over all the scatter plots. Notably, while the total count indicates a preference

for a line approaching the vertical regression line, this is not the case for each individual scatter plot.

Table 2
A count of how many participants had a smaller vertical, horizontal, or orthogonal L-2 norm for each scatter plot.

Number of Users with The Smallest Vertical, Horizontal, Orthogonal l-2 norms			
Scatter Plots	Vertical	Horizontal	Orthogonal
1	39	47	17
2	101	1	1
3	9	55	39
4	49	17	37
5	26	15	62
6	80	0	23
7	19	37	47
8	63	26	13
9	14	37	52
All	400	235	291

When, in Method 1, the Chi-Square Homogeneity Test is applied to every scatter plot for all categories at once, the null hypothesis is rejected for every scatter plot as shown in Table 3. This means that for no scatter plot are all three regression lines equally preferred. When applied to each matched pair, the Chi Square Homogeneity Test demonstrates that in some instances there is a lack of sufficient evidence to reject the claim that there is not a preference, but most of the time, there was a preference between the pair.

Table 3
The results from the Chi-Square Homogeneity Test.

Chi Square Homogeneity Test				
		Pairs		
Scatter Plot	All Categories	*H-V*	*H-O*	*V-O*
1	Reject	Fail to Reject	Reject	Reject
2	Reject	Reject	Fail to Reject	Reject
3	Reject	Reject	Reject	Reject
4	Reject	Reject	Reject	Fail to Reject
5	Reject	Fail to Reject	Reject	Reject
6	Reject	Reject	Reject	Reject
7	Reject	Reject	Fail to Reject	Reject
8	Reject	Reject	Fail to Reject	Reject

9	Reject	Reject	Reject	Reject

Table 4
The calculated 95% confidence interval of the average mean difference for each matched pair.

95% Confidence Interval of the Average Means for the Matched Pairs						
	H-V		*H-O*		*V-O*	
Scatter Plot	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
1	-0.0942	0.0487	-0.01266	0.0518	0.0057	0.0803
2	730.1146	777.9488	653.1303	693.4867	-85.3779	-77.4433
3	-0.8412	-0.0516	0.3174	0.8217	0.8364	1.1896
4	-3.9077	-2.4922	-2.3563	2.5949	0.8886	5.8583
5	0.3491	1.3046	0.5670	0.9720	-0.3846	0.2627
6	0.4990	1.6046	1.1664	1.8526	0.2286	0.7033
7	5.0900	5.8506	3.4659	3.6753	-2.2100	-1.6004
8	-0.6243	-0.1341	-0.0146	0.3304	0.4517	0.6372
9	0.1194	0.3651	0.0907	0.1661	-0.1973	-0.0269

Table 4 shows that the average 95% confidence interval of the mean difference between each matched pair can provide evidence that a fitted line is closer to one regression line than another. For example, by looking at the confidence interval for the first matched pair for the second scatter plot, the lower and upper bounds are 730.1146 and 777.9488 respectfully. Since the average distance, which is taken from the bootstrapping method, is far above 0, it can be assumed that the average distance from the fitted lines to the horizontal regression line is much larger than the average distance to the vertical regression line. This would clearly indicate that fitted lines are closer to the vertical regression line than the horizontal regression line for this scatter plot.

In the following report of findings, three graphs are provided for each of the data sets. First, the scatter plot of the data set is provided (Figures "*X*" left panel). On the scatter plot, four graphs are provided: the vertical regression line (red); the horizontal regression line (blue); the orthogonal regression line (cyan); and the average of the participants' lines of best fit (black). This plot will show the geometric relationship between the fitted line and the regression lines. Second, the scatter plot of participants' results is provided in respect to their lines and each linear function is coded as an ordered pair (*slope*, *intercept*); additionally, plotted are ordered pairs representing each of the vertical, horizontal, and orthogonal regression lines (Method 1) (Figures "*X*" centre panel). This plot will show if a relationship holds between the slope and intercepts of the fitted lines. Third, the scatter plot of participants' results is provided in respect to their lines in the form of slope and distance from the centroid of the data. Each linear function is coded as an ordered pair (*slope*, *distance*) and all these ordered pairs are plotted along with the ordered pairs representing each of the vertical, horizontal, and orthogonal regression lines (research Method 2) (Figures "*X*" right panel). Since the distance from each regression line to the centroid

is 0, the ordered pairs for the regression lines will always be plotted on the horizontal axis. This plot will show the variation in the slopes and intercepts of the fitted lines and the position of the slopes and intercepts of the regression lines relative to the boxplots.

The results of three of the nine scatter plots are presented below in detail with concluding summary statements. The remaining scatter plots will be discussed singularly through parallel summary statements.

Scatter Plot 1.

According to Method 1, the results for Scatter Plot 1 show that more lines fitted by participants are closer to the horizontal regression line than the vertical and orthogonal regression line (Table 2). However, from the Chi Square Homogeneity Test (CSHT) for *H-V* (Table 3), there isn't enough evidence to support the claim that lines fitted by participants have a preference for being closer to the horizontal regression line than the vertical regression line. When considering other matched pairs (*H-O* and *V-O*), the CSHT reveals that there is enough evidence to support the claim that there is not an equal likelihood of a line fitted by participants being closer to the orthogonal regression line than the horizontal regression line, as well as the vertical regression line. In respect to the average 95% confidence interval for the mean (Table 4):

- The range [-0.0942, 0.0487] for *H-V* indicates that there is no preference to which regression line the fitted lines are closer;
- The range [-0.0127, 0.0518] for *H-O* suggests that the average distance between the fitted lines and the orthogonal regression line is less than the average distance between the fitted lines and the horizontal regression line; and
- The results for *V-O* show a lower and an upper bound that are both positive with a small difference.

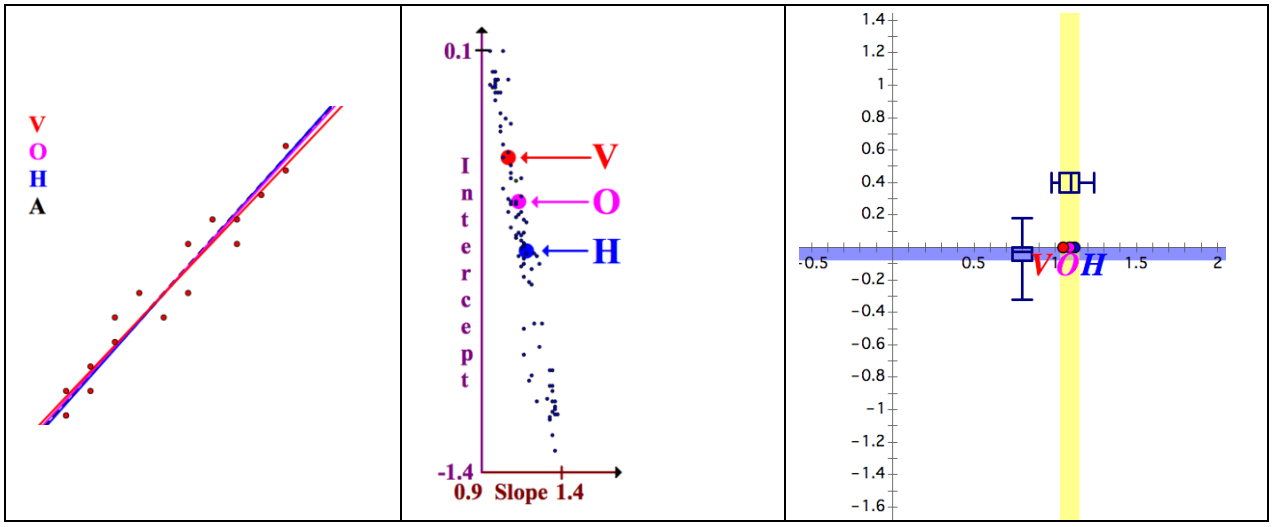

Figure 2. Scatter Plot 1 superimposed with vertical, orthogonal, and horizontal regression lines and the participant average line of fit (left); a scatter plot of the slope and intercept of participants' lines of fit, with the slope and intercept points of the three regression lines (centre) (Method 1); and boxplots of participants' lines of fit (the interquartile range of the slope is in yellow and the distance from the data's centroid is in purple) with regression lines being reported as slope and centroid (right) (Method 2).

Therefore, even though there are more fitted lines closer to the vertical regression line than orthogonal regression line, on average, the area between the fitted lines and the orthogonal regression line is smaller. While the fitted lines were closer to the horizontal regression line than the other regression lines, they were also closer to the orthogonal regression line than the vertical regression line. However, the tight linearity of the data and the proximity of the three regression lines make it difficult for an analysis of the comparative distances between the participants' average line and the regression lines.

In summary, for Scatter Plot 1, results indicate that even though a greater number of lines were fitted closer to the horizontal regression line than the other regression lines, there isn't enough evidence to suggest that there is preference for fitted lines being closer to the horizontal regression line than the vertical regression line.

Method 2 demonstrates that the interquartile range of the slopes of the participant lines (in yellow) are tightly associated with the slopes of the three regression lines and that the interquartile range and of the distance from the centroid lies close to but slightly below the centroid of the data. Summarily, with this highly linear data, with no outliers, while the participant's lines of fit are quite close to each of the regression lines, they are also quite tightly compacted with each other with very little variation.

Scatter Plot 2.

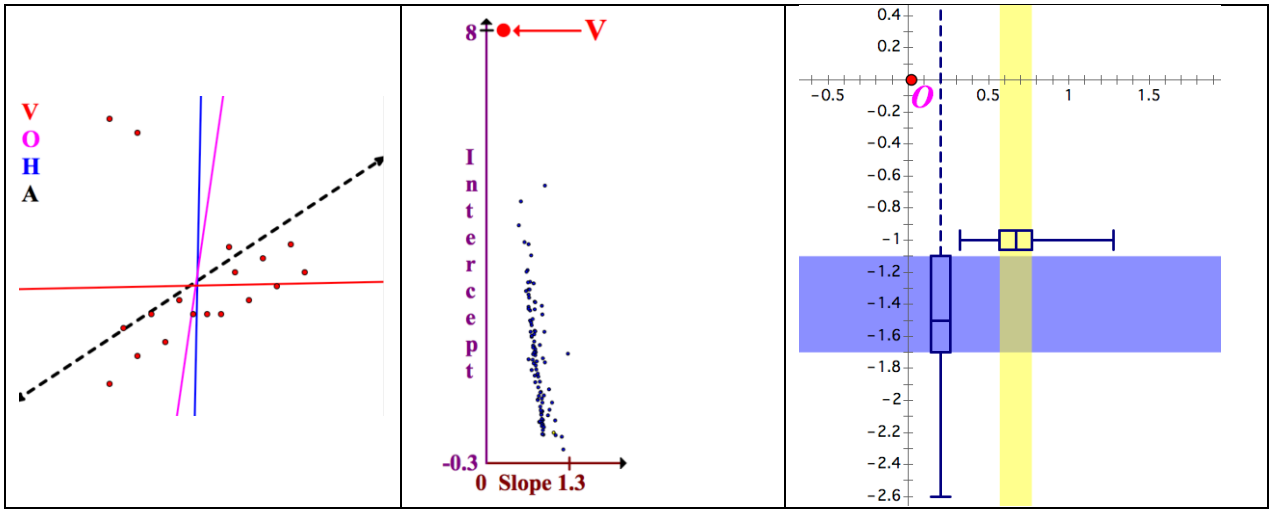

Figure 3. Scatter Plot 2 and graphs similar to Figure 2.

Data set 2 has most points showing a linear trend, but with two outliers far above the general trend. Based on Method 1, regarding Scatter Plot 2, Table 1 shows that all – except two – lines fitted by participants are closer to the vertical regression line. The CSHT (Table 3) shows that fitted lines are not equally likely to occur nearest the three regression lines. It is recommended to reject the null hypothesis for *H-V* and *V-O*, meaning that fitted lines do have a preference for being closer to the vertical regression line than the other regression lines. However, for *H-O* it was recommended to fail to reject the null hypothesis, meaning that there is insufficient evidence to claim that lines fitted by participants are closer to the orthogonal regression line than the horizontal regression line. In respect to the average 95% confidence interval of the means (Table 4):

- For *H-V*, the large positive lower and upper bounds indicate that the average distance between the fitted lines and the horizontal regression line is greater than the distance between the fitted lines and the vertical regression line;
- For *H-O*, where both the lower and upper bounds are large positive values, on average, the distance between the fitted lines and the horizontal regression line is larger than the average distance between the fitted lines and the orthogonal regression line. Even though the CSHT failed to reject the null hypothesis for *H-O*, it has to be taken into account that a majority of lines fitted by participants are closer to vertical regression line. Since only one fitted line was closest to the horizontal and orthogonal, the CSHT test couldn't detect a difference; however, the confidence interval clearly indicates the fitted lines were closer to the orthogonal regression line.
- The range [-85.3779, -77.4433] for *V-O* suggests that the average distance between the vertical regression line and the fitted lines is less than the distances between the orthogonal regression line and the fitted lines.

Notably, the right-hand scatterplot comparing participant generated lines with the vertical, orthogonal, and horizontal regression line does not have points representing the horizontal and orthogonal regression lines. This is due to scale. The orthogonal line would be plotted at (6.8, -41.7) and (51.5, -367.7), both far beyond the scale of the presented axes. Thus, it is clear that almost all participant lines of best fit are closer to the vertical regression line than the others. Summarising results for Scatter Plot 2, lines fitted by participants were closer to the vertical regression line than the other two regression lines.

Method 2 demonstrates that outliers can significantly affect participants' placements of lines. The boxplots reveal a significantly larger range (in total and in the interquartile range) for both the slopes and the distances of the participants' lines from the centroid. Thus, the inclusion of outliers causes more variability in participant results – even to the extent where participants' lines are generally quite apart from any of the three regression lines. However, this finding may be somewhat misleading. Noting that outliers can greatly impact the location of the regression lines, it can be argued that none of the regression lines well represent the data – with the vertical and orthogonal lines being particularly poor. In fact, from an intuitive perspective, it can be argued that the participants' average line, better modelled the trend of the data than the regression lines.

While there is greater variability in the participants' lines of fit, the interquartile ranges of the slopes and distances from the centroid remain relatively compact. This may imply that participants employed a relatively consistent pattern recognition method to handle outliers.

Scatter Plot 3.

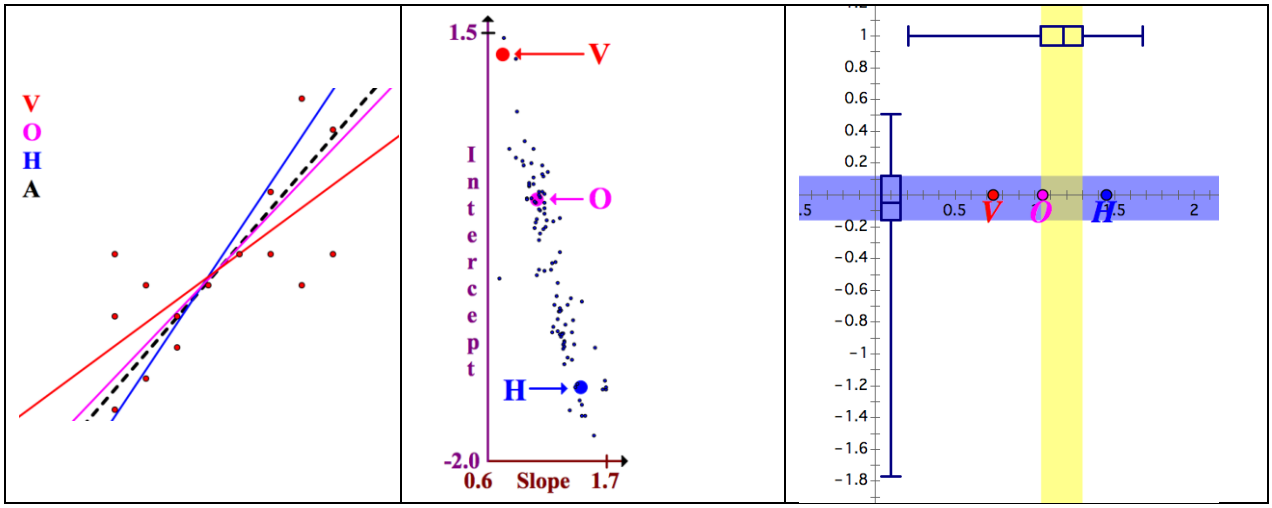

Figure 4. Scatter Plot 3 and graphs similar to Figure 2.

This scatter plot corresponds to data set #3 in the charts and this data has a weak linear trend. Based on Method 1, Scatter Plot 3 is one of the few scatter plots to have more fitted lines being closer to the orthogonal regression line than the vertical and horizontal regression lines (Table 2). When applying the CSHT to each of the matched pair (Table 3), the null hypothesis is rejected, meaning that fitted lines have a preference towards one regression line than the other regression line for each matched pair. Regarding the 95% confidence interval of the average means (Table 4):

- The range [-0.8412, -0.0516] for *H-V* indicates that the average distance between the fitted lines and the horizontal regression line is smaller than the average distance of the area between the fitted lines and the vertical regression line.
- The range [0.3174, 0.8217] for *H-O* indicates that the average distance between the fitted lines and the orthogonal regression line is smaller than average distance between the fitted lines and the horizontal regression line.
- The bounds for *V-O* ([0.8364, 1.1896]) show that the average distance between the fitted lines and the vertical regression line is larger than the average distance between the fitted lines and the orthogonal regression line. Once again, it has to be taken into consideration that only the vertical and orthogonal L-2 norms are being considered. When only considering the area between the fitted lines and the vertical regression line and the orthogonal regression line, participants overall have smaller average area between the user's line and the orthogonal regression line than the average area between the user's line and the vertical regression line.

The right-hand scatter plot demonstrates that a greater number of participant generated lines cluster closer to the orthogonal regression line than the others. This is more readily seen in this case than in Scatter Plot 1. Overall, for Scatter Plot 3, fitted lines are closer to the orthogonal regression than the horizontal and vertical regression lines.

Intuitively, while one may induce a general linear form to the data, this data set does not contain notable outliers. According to Method 2, the interquartile range of the participants' slopes of their lines is quite compact and clustered closer to the orthogonal regression line than

the other regression lines and the interquartile range of the distance of the participants' lines from the centroid is comfortably centred on the centroid of the data. Thus, unlike for Scatter Plot 1, here there is a balance of participants' lines above and below the centroid.

Scatter Plot 4.

For Scatter Plot 4, more lines fitted by participants are statistically closer to the vertical regression line than the horizontal and orthogonal regression lines, as shown in Table 2. However, as shown in the accompanying graphs, the proximity of the regression lines causes this preference to be marginal, as demonstrated in the distribution of participants' lines shown in the centre panel of Figure 5. In conclusion, the results show that more lines fitted by participants are closer to the vertical regression line; however, it cannot be concluded that more fitted lines would be fitted closer to the vertical regression line over the orthogonal regression line.

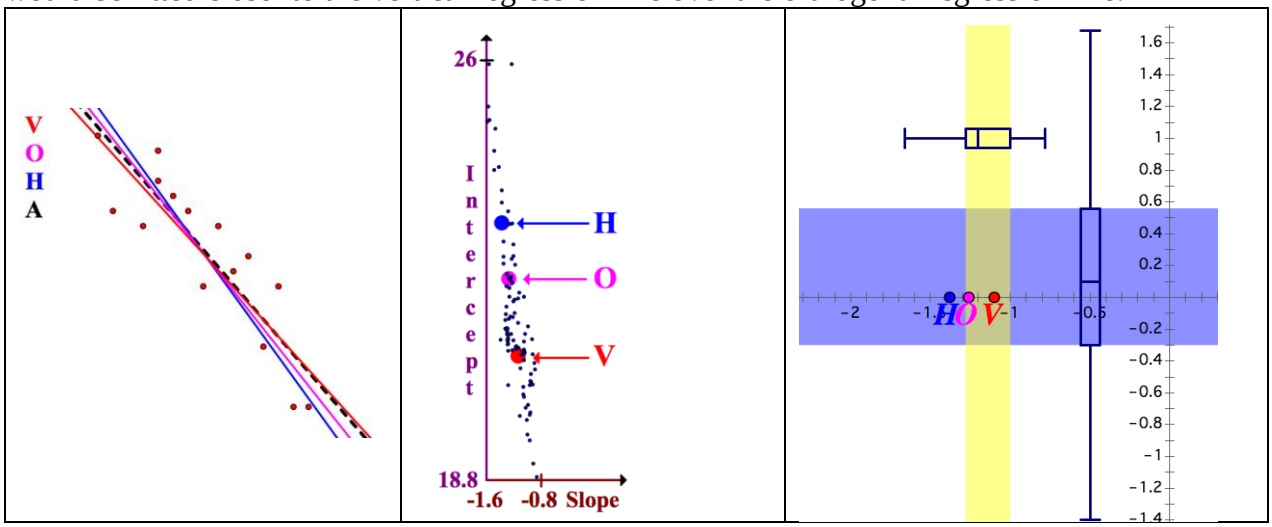

Figure 5. Scatter Plot 4 and graphs similar to Figure 2.

As revealed through Method 2, the width of the participants' interquartile ranges in both slope and distance from the centroid is significantly larger in respect to this data set than for others. This is particularly interesting in that the data set is relatively linear in nature with no notable outliers. In fact, the primary distinction between this data set and previous data sets is that the regression lines to this scatter plot have negative trends.

Scatter Plot 5.

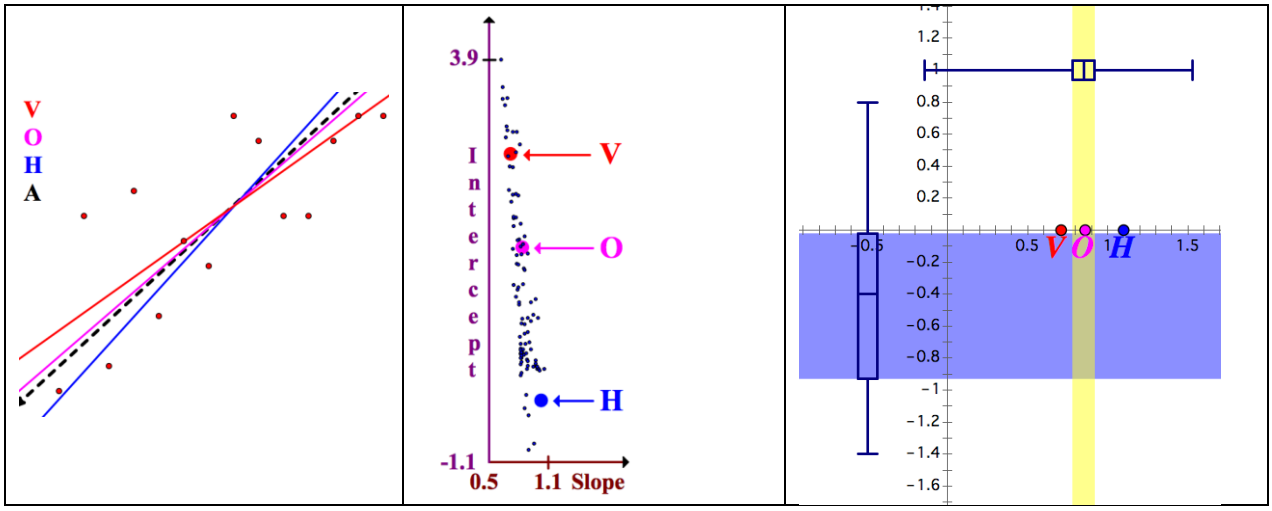

Figure 6. Scatter Plot 5 and graphs similar to Figure 2.

Scatter Plot 5 has more fitted lines being statistically closer to the orthogonal regression line. However, the scatter plot demonstrates a significant clustering of participant lines between the orthogonal and horizontal line. Nevertheless, these results show that lines fitted by users are more likely to be closer to the orthogonal regression line than the other regression lines.

Noting that the correlation coefficient for the linear regression for this data set is quite low, Method 2 reveals interesting participant behaviour. While the interquartile range of the slopes of the participants' lines is quite compact and closer to that of the orthogonal regression line, the interquartile range of the lines' distances from the centroid is quite wide and most participant lines were positioned below the centroid. Altogether, this may imply that participants are better at matching slopes in pattern recognition than they are at placing their lines a particular distance from the centroid when the data is less linear.

Scatter Plot 6.

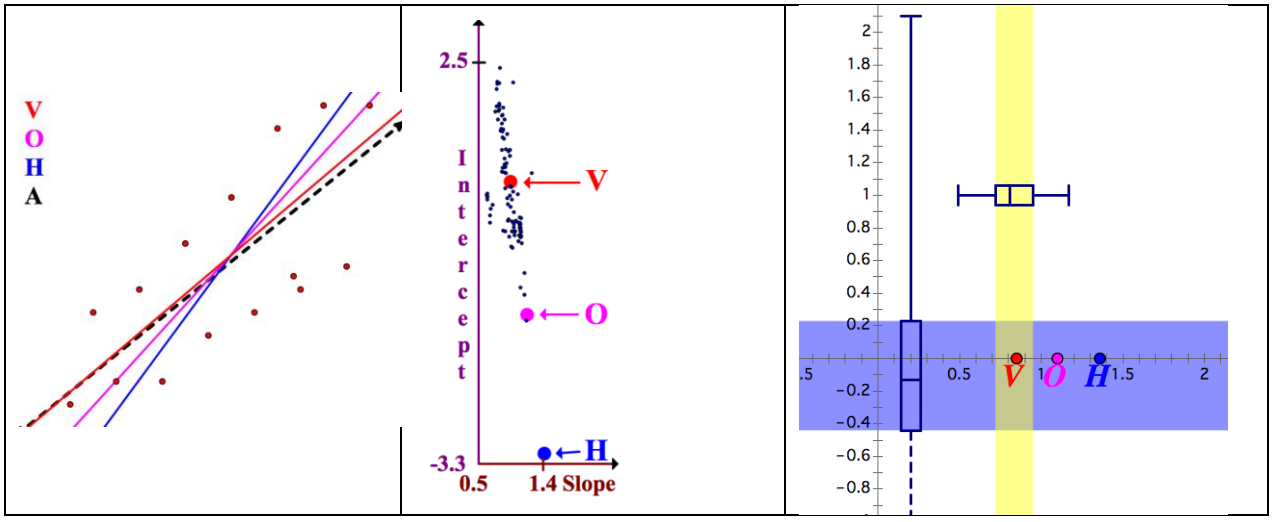

Figure 7. Scatter Plot 6 and graphs similar to Figure 2.

For Scatter Plot 6, an overwhelming number lines fitted by participants were closer to the vertical regression line than the horizontal and orthogonal regression lines. This is readily apparent looking at the numbers in Table 2 and at the scatter plot in the middle.

As in the previous case, Method 2 reveals that, when the correlation coefficient of the data is rather low, the slopes of participants' lines remains more compact than the distances of the lines to the centroid and, in this case, the slopes of the participant line trend toward the slope of the vertical regression line.

Scatter Plot 7.

The statistical results from Scatter Plot 7 show that there is a somewhat even distribution among the lines fitted closer to the orthogonal and horizontal categories, while few fitted lines were close to the vertical regression line. However, the right-hand scatter plot may reveal a preference to the orthogonal line. In conclusion, even though more fitted lines were closer to the orthogonal regression line than the horizontal regression line, there isn't enough evidence to suggest that fitted lines would be closer to the orthogonal regression line than the horizontal regression line.

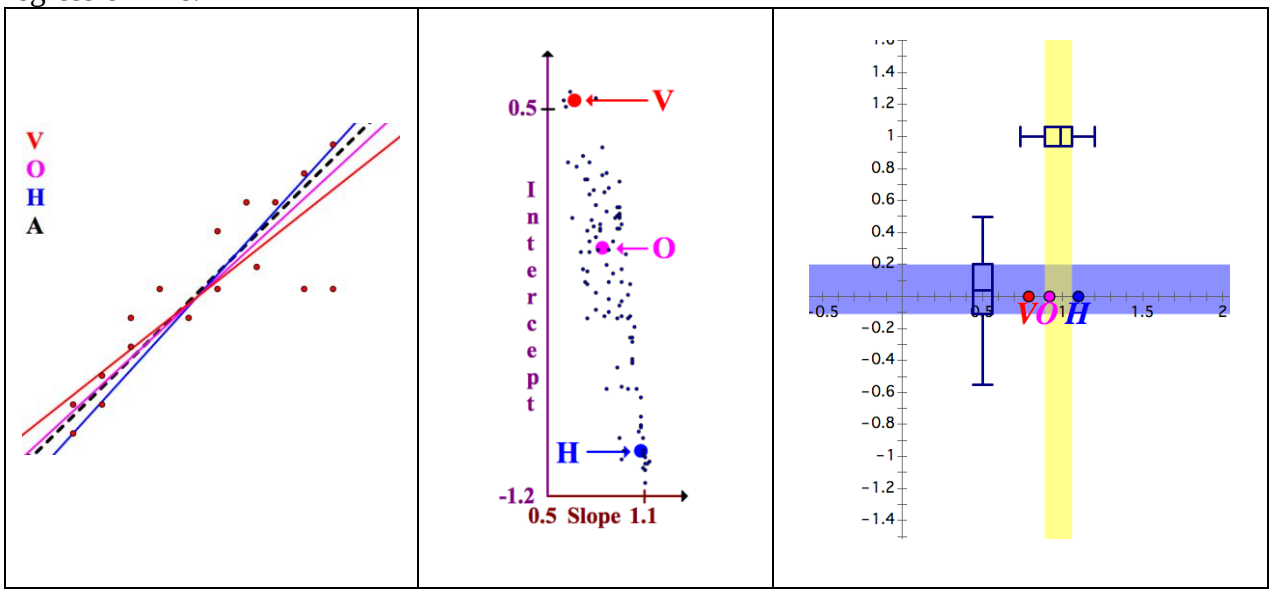

Figure 8. Scatter Plot 7 and graphs similar to Figure 2.

In respect to Method 2, this case begs the question: how disjoint from the data must an outlier be to be treated differently by participants? In this case, it seems as if the two potential outliers did not disturb the participants' perception of the pattern of the data. The interquartile ranges of both the slopes and the distances are quite compact. Thus, participants see the "pattern" of the data quite consistently.

Scatter Plot 8.

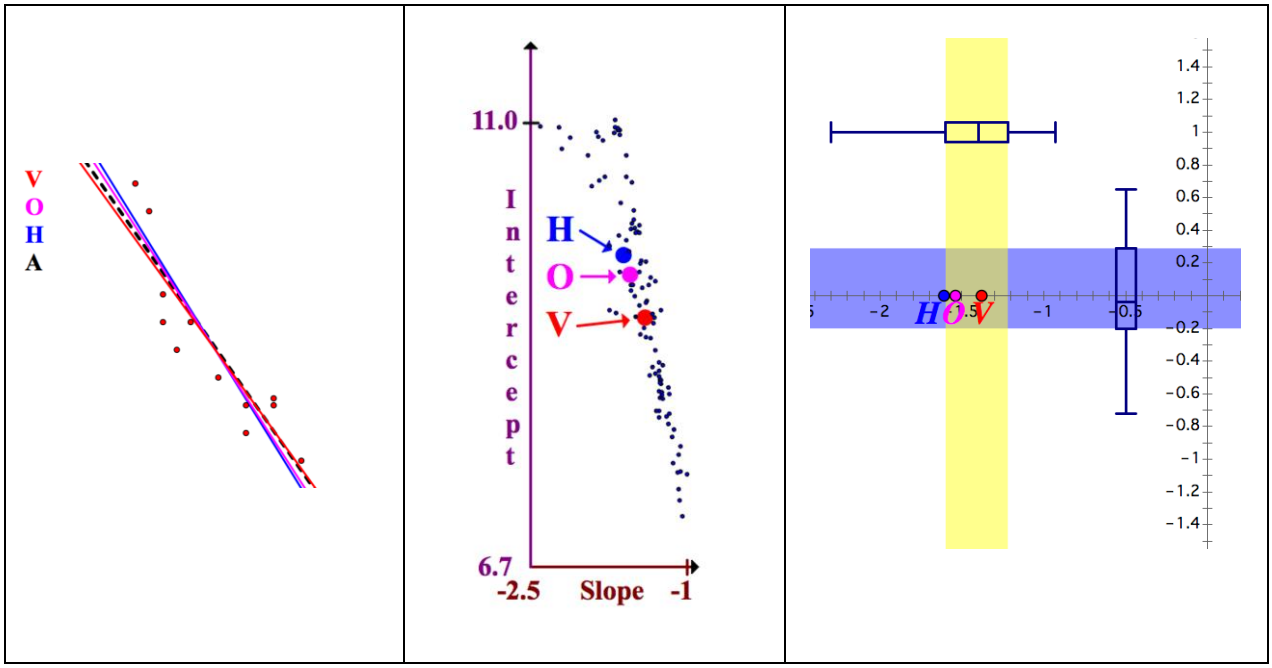

Figure 9. Scatter Plot 8 and graphs similar to Figure 2.

The counts from Scatter Plot 8 show that a majority of fitted of lines are closer to the vertical regression line than the other regression lines. The right-hand scatter plot shows many participant lines have a slope larger than the horizontal or orthogonal regression lines, while were evenly distributed above and below the slope of the vertical regression line.

As in Scatter Plot 4, this data has a negative trend with, again, no notable outliers. Herein, Method 2 reveals that the width of the participants' interquartile range on the slope is greater than for many other data sets. However, the width of the interquartile range of the distances from the centroid is quite compact and relatively centred on the centroid, significantly larger in respect to this data set than for others. It must be again wondered if participants see patterns differently if the data trends upward rather than downward.

Scatter Plot 9.

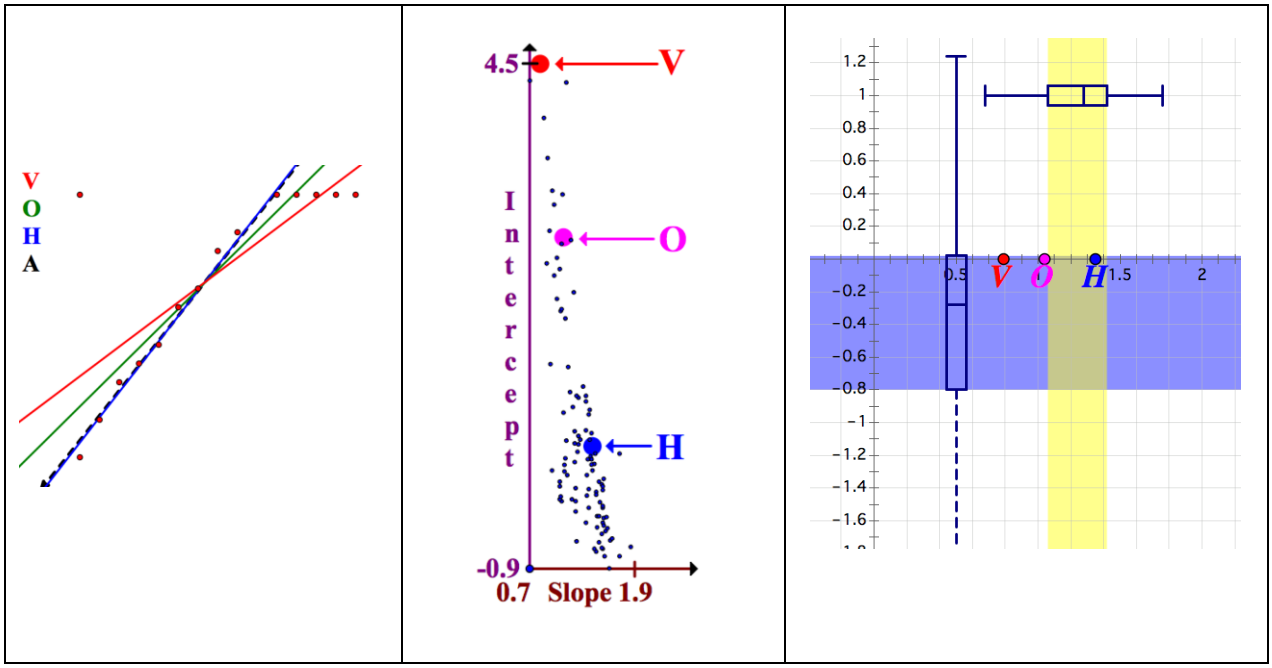

Figure 10. Scatter Plot 9 and graphs similar to Figure 2.

For Scatter Plot 9, the majority of fitted lines tend to be closer to the horizontal regression line to the other regression lines. Notice that for this data set, the horizontal regression line follows the general linear trend of the majority of the data and doesn't seem to be impacted as much by the points not along this trend. The participants seem to fit lines reflecting this linear trend and also seem less affected by the outliers.

In this plot, one could again question which regression line best represented the trend of the data. Nevertheless, the interquartile range of the distances of the participants' lines from the centroid is quite wide and clearly tends to be below the centroid.

Discussion

As shown in Table 1, there were more lines fitted by participants that are closer to the vertical regression line than the horizontal and orthogonal regression lines throughout the entire experiment. However, it should not be assumed that participants are overall more likely to construct lines of best fit closer to the vertical regression line than the other lines without considering the shape and trend of the data as well as any outliers.

Several observations are made by considering either several scatter plots together rather than separately or all scatter plots together. First, as seen by a comparison of Scatter Plots 1 and 2, outliers in a distribution affect how many fitted lines are closer to one regression than the others. Scatter Plot 1 has a strong linear correlation and a variation in how many fitted lines were closer to each regression line. Scatter Plot 2 also has a strong linear correlation but has two outliers. These outliers significantly affected the regression equations and seemed to increase the number of fitted lines closer to the vertical regression line than the other regression lines. (Caveats to this finding are discussed later.) However, and contradictorily, most participants seemed to ignore or omit the outliers when fitting a line to Scatter Plots 2 and 8.

The literature would seem to suggest that, in some occasions, outliers perturb user interpretation of data trends. Outliers can be confounding facts in a representation, particularly when the user interacts syntactically with the data (considering smaller facets of the data rather than seeing the data globally). However, in some instances, outliers are seemingly simply overlooked or avoided. This is consistent with the literature stating that, when mathematical representations are interpreted, students often fail to discern between valuable information and what can be ignored (e.g., Bossé, Adu-Gyamfi, & Chandler, 2014; Duval, 2006; Kaput, 1987b; Lesh, Post, & Behr, 1987; and Sternberg, 1984). Concatenating these ideas, it seems that, at times, outliers may have confounded participant interpretation of the data and overly perturbed their visualised lines of best fit in one way or another and that, at other times, the significance and possible effect of these outliers on the visualised lines may have been overlooked.

The different trends and patterns various scatter plots in this experiment shows that lines fitted by the participants have a preference to be closer to one regression line than another regression line for scatter plots with a moderate linear trend. However, there seems to be a preference to one regression line or another when lines are fitted through scatter plots in which there is a variation and departure from a linear trend. When participants fitted lines to a scatter plot with an exponential pattern (i.e., Scatter Plot 8), a majority of fitted lines were closer to the horizontal regression line than the other regression lines. For Scatter Plot 3 representing a piecewise pattern involving two linear trends, a majority of lines fitted by participants are closer to the orthogonal regression line than the other two regression lines. Summarily, for nonlinear trends on scatter plots, there seems to be variation regarding which regression line most fitted lines approximated, and this preference seems to be affected by the types of patterns represented in the scatter plots.

It was previously noted that outliers may act as confounding characteristics in a scatter plot. So, too, may distributions that have poorer linear correlation act in a confounding manner. However, it may not be this simple. While it is sensible that scatter plots with stronger linear trends are interpreted more consistently than distributions with greater variation, it cannot be determined with certainty whether this is due to an inherently confounding nature. Furthermore, it may be that distributions with poorer linear correlations are more prone to syntactic interpretations, and distributions with higher linear correlations are more prone to semantic interpretations, as the latter can more readily be perceived as a global, well-patterned whole.

Comparing results from Scatter Plots 1 and 4, the Chi Square Homogeneity test shows evidence that for moderate to strong linear trends, the likelihood of lines fitted closer to the most often associated regression line is not greater than the likelihood of fitted lines being closer to the second most often associated regression line. For moderate to strong linear trends, there isn't enough evidence to suggest that when participants fit their line of best fit through these scatter plots that their lines would be closer overall to one regression line.

Method 2 focuses on the variation in the participant's lines by graphing the spread of the slopes, which indicates the direction of the line, and orthogonal distances from the centroid, which indicates a translational closeness of the fitted line to a given regression line. For scatter plots with little variation and positive slope (Scatter Plots 1, 2, and 7), there was little variation in the slopes of the fitted lines. Even with the presence of outliers (2 and 7), the slopes of the participants fitted lines were very similar. The distances from the centroid showed different results. There was more variation in the distances from the fitted lines on Scatter Plots 2 and 7, containing outliers, than in Scatter Plot 1 where there were no outliers. Scatter Plots 3, 5, and 6 show a weaker positive relationship and there is a larger variation in the distances from the

centroid. However, the middle 50% of the slopes for these scatter plots show little variation, but there is a large spread in the tails. For scatter plots with a positive trend, participants are consistent with the directions of their fitted lines, but less consistent with their vertical location. Scatter plots 4 and 8 have a negative trend, and even though the relationship is strong with no real outliers, participants were less consistent with their slopes and distances from the centroid than with data sets with a positive relationship.

Modified Results

Notably, the results from Scatter Plot 2 significantly skew the results of the total experiment. Almost all participants constructed their lines of fit most closely to the vertical regression line and well away from the other two regression lines. In comparison to the other scatter plots, the results from Scatter Plot 2 can be interpreted as an anomaly. Thus, it was questioned what would be the results for the total experiment if Scatter Plot 2 was omitted. Table 5 provides these comparative results.

Table 5
The Chi-Square Homogeneity Test results when applied to the entire experiment with and out the second scatter plot data.

Chi-Square Homogeneity Test	Pairs			
Keep All Results or Omit Plot 2	All Categories	H-V	H-O	V-O
All Results	Reject	Reject	Reject	Reject
Omit Plot 2	Reject	Reject	Reject	Fail to Reject

The only difference is that, without Scatter Plot 2, there is insufficient evidence to reject the statement that a participant's fitted line would be equally likely to be closer to the vertical versus the orthogonal regression lines. It is possible that a participant's line is closer to the vertical and orthogonal with equal probability.

Unanticipated Findings

Notably, all vertical, orthogonal, and vertical regression lines pass through $(\overline{x},\overline{y})$, the centroid of the data. Interestingly, the line representing the mean of the lines of best fit for each scatter plot also intersects the three regression lines very near the centroid. This unanticipated finding is recognisable through the graphs of the regression lines throughout this paper (it as not yet been statistically analysed) and brings rise to an important question regarding human pattern recognition in respect to linear trends in scatter plots. Do people intuitively recognise the centroid of a set of data even more so than a linear trend through the data? If participants better recognise data centroids, it is sensible that, as demonstrated throughout results herein, the slope of their line of fit would have less variability than the line's *y*-intercept. Even a minor alteration in the slope of a line can produce a significant change in the line's *y*-intercept. (We hope to investigate this in more detail in future studies.)

In respect to representational interpretation, it seems that users more readily interpret the centroid of the data – with less confounding potential – than any other characteristic of the

distribution. There is possibility that this characteristic is most semantically interpreted with the slope of the line of fit only slightly less so. However, the intercept of the line of best fit seems most syntactically interpreted.

Method 2 showed little variation in the slopes of the participant's fitted lines for data sets with a medium to strong positive linear relationship, even with a few outliers. However, there was much more variation in the slopes for data sets with a medium to strong negative linear relationship, even without outliers. It was surprising that participants seemed to be more uncertain about the angle of the trend in negatively correlated data. Also, there was more variation in the distance from a participant's fitted line to the centroid of the data, and this effect was amplified by the presence of outliers. It would be interesting to see if the factors that affect the vertical positioning of the fitted lines can be identified.

This, again, speaks to the potential confounding nature of the slope of the perceived line of best fit. The fact that participants possibly interpreted data trends with positive correlations more semantically and data trends with negative correlations more syntactically, is both illuminating and difficult to fully understand based singularly on this investigation. Nevertheless, positive versus negative correlated data certainly seems to confound many.

Summary and Implications

Pattern recognition is the ability to process, interpret information, and detect patterns about an environment. The results of this study show how participants analyse different patterns and fit a line of best fit, seeing how these participants might be able to predict different values not included in each data set. Participants react differently to different patterns and trends, fitting lines closely to one regression line to what they believe is the line of best fit. The variation of which regression line participant's line of best fit shows how each individual interprets patterns differently. This type of variation isn't shown when using computers, as computers have difficulties detecting trends, patterns, or anomalies in visualised data. As technology evolves in displaying data visually, human pattern recognition will also evolve, improving the ability of humans to detect patterns in new areas.

Scatter plots used in this study exhibited linear, exponential, cluster, and piecewise trends in the data and some of these scatter plots contained outliers. In several cases, it would be more appropriate to fit a curve instead of a straight line to model the data. Nevertheless, the task was to place a line of best fit through the data. (Left uninvestigated was how well users could have fit curves to data when appropriate.) Throughout the study, it can be seen that there were more fitted lines closer to the vertical regression line than the horizontal and orthogonal regression lines. However, this finding was contingent upon the trend of the data in particular scatter plots. Additionally, outliers had much less impact on lines fitted by participants than had been hypothesised.

A number of practical concerns for classroom teachers arise from this investigation. First, we recognise that numerous important concepts are addressed in the Common Core standards, including: deviations from the overall pattern of the data (outliers); positive or negative association; linear and nonlinear associations; informally and formally fitting a straight line to the data; informally and formally assessing the model fit of linear and other appropriate functions to the data. This study shows that not all concepts are equal – or, rather, some concepts may be more difficult to instruct and assess than others. Teachers must well recognise that some of these topics can significantly affect the learning, understanding, and assessing of other concepts.

Second, teachers must recognise that an issue as seemingly trivial as the positive or negative trend of data might affect students' mathematical understanding. While seen from this study that the positive or negative trend of the data affects student's positioning of lines of fit through scatter plots, it may as well be that students understand linear functions with positive slopes more than negative slopes. (This would necessitate additional research.) Thus, any finding in this study may affect a far wider swath of mathematics.

Third, the classroom teacher must carefully select problems for students to investigate. As not all concepts are the same, neither are all problems. Thus, the importance of the proper selection of problems cannot be overstated. One problem may be more fraught with confounding information than another. Unfortunately, more fully explicating precise types of problems that should first be investigated and those that should be delayed cannot, herein, be sufficiently explicated. Indeed, this requires far more investigation.

Finally, the classroom teacher can use the curious nature of scatter plots, lines of fit, and the findings in this investigation as inducements to gain student engagement in problems involving lines of fit. Rather than dismissing the fact that students produce different lines of fit to particular scatter plots, this can be rich fodder for students to further investigate these lines.

While altogether, this study demonstrates that student lines of best fit cannot be simply recognised as good (approximating the vertical least squares regression line) or poor (departing significantly from the same line), a number of implications for education are immediately apparent from this study. First, while it may seem that a scatter plot is simply a scatter plot, there are clearly characteristics associated with some scatter plots (e.g., the existence of outliers, strong or weak linear correlation, and whether the correlation is positive or negative) that affect students' interpretations and interactions with them. Educators must recognise that within any mathematical representation there may exist confounding attributes.

Second, educators must recognise that students often perceive a particular problem in different ways. Some may interact semantically and others syntactically with the representation. This may gravely affect student mathematical performance. Indeed, mathematical success or failure on a particular problem may not be an indication of understanding as much as representational interpretive style.

The third implication may be the most profound. This study clearly demonstrates the need for further investigation. Since this group of participants was somewhat homogeneous, it must be wondered if students in other fields of study with significantly different backgrounds and experiences with data and the associated fitting of lines would have had similar or different results. Extending findings beyond scatter plots, it would be valuable to investigate any number of other mathematical representations (e.g., function graphs, function tables, diagrammatic representations, histograms, and countless other representations) for confounding attributes and student preferences for symmetric or semantic interaction. It is hoped that future research can expand upon these findings, taking into account this study's limitations, and provide new insight in how humans recognise patterns and visualise lines of best fit and interact with mathematical representations.

References

Adu-Gyamfi, K. & Bossé, M. J. (2014). Processes and reasoning in representations of linear functions. *International Journal of Science and Mathematics Education, 12*(1), 167-192.

Adu-Gyamfi, K., Bossé, M. J., & Chandler, K. (2015). Situating student errors: Linguistic-to-algebra translation errors. *International Journal for Mathematics Teaching and Learning*. Retrieved from http://www.cimt.plymouth.ac.uk/journal/bosse6.pdf.

Adu-Gyamfi, K., Stiff, L., & Bossé, M. J. (2012). Lost in translation: Examining translation errors associated with mathematics representations. *School Science and Mathematics, 112*(3), 159-170.

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education,33*(2-3), 131-152.

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics 52*(3), 215-241.

Bossé, M. J., Adu-Gyamfi, K., & Chandler, K. (2014). Students' differentiated translation processes. *International Journal for Mathematics Teaching and Learning*. Retrieved from http://www.cimt.plymouth.ac.uk/journal/bosse5.pdf.

Bossé, M. J., Adu-Gyamfi, K., & Cheetham, M. (2011). Assessing the difficulty of mathematical translations: Synthesizing the literature and novel findings. *International Electronic Journal of Mathematics Education, 6*(3), 113-133.

Duval, R. (2006). A cognitive analysis of problems of comprehension in the learning of mathematics. *Educational Studies in Mathematics, 61*, 103-131.

Embse, C. V. (1997). Visualizing least-square lines of best fit. *The Mathematics Teacher 90*(5), 404-408.

Goldin, G. A. (1987). Cognitive representational systems for mathematical problem solving. In Claude Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 125-145). Hillsdale, New Jersey: Erlbaum.

Janvier, C. (1987). Translation process in mathematics education. In C. Janvier (Ed.), *Problems of representations in the teaching and learning of mathematics* (pp. 27-31). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kaput, J. J. (1987a). Representation systems and mathematics. In C. Janvier (Ed.), *Problems of representation in teaching and learning mathematics* (pp. 19–26). Hillsdale, NJ: Erlbaum.

Kaput, J. J. (1987b). Toward a theory of symbol use in mathematics. In C. Janvier (Ed.), *Problems of representation in mathematics learning and problem solving* (pp. 159-195). Hillsdale, NJ: Erlbaum.

Leng, L., Zhang, T., Kleinman, L., & Zhu, W. (2007). Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. *Journal of Physics: Conference Series 78* 012084, doi:10.1088/1742-6596/78/1/012084. Retrieved from http://iopscience.iop.org/article/10.1088/1742-6596/78/1/012084/pdf.

Lesh, R., Post, T., & Behr, M. (1987). Representations and translations among representations in mathematics learning and problem solving. In C. Janvier (Ed.), *Problems of representations in the teaching and learning of mathematics* (pp. 33–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Moore, D. S., Notz, W. I., & Fligner, M. A. (2015). *The basic practice of statistics, Instructor's edition* (7th ed.). New York: W. H. Freeman.

Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981). Eye fitting straight lines. *The American Statistician* 35, 150-152.

Motulsky, H. J., & Ransnas, L. A. (1987). Fitting curves to data using nonlinear regression: A practical and nonmathematical review. *The FASEB Journal 1*(5), 365-374.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common core state standards in mathematics*. Washington D.C.: Authors.

Olson, D., & Campbell, R. (1993). Constructing representations. In C. Pratt & A. F. Garton (Eds.), *Systems of representations in children: Development and use* (pp. 11-26). New York, NY: John Wiley & Sons, Inc.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Steinbring, H. (1997), Epistemological investigation of classroom interaction in elementary mathematics teaching, *Educational Studies in Mathematics, 32*, 49-92.

Sternberg, R. J. (1984). Mechanisms of cognitive development: A componential approach. In R. J. Stemberg (Ed.), *Mechanisms of cognitive development* (pp. 163-186). New York: W. H. Freeman.