On the Usefulness and Limitations of Diagrams in Statistical Training

Atsushi Terao (atsushi@edu.hokudai.ac.jp)
Graduate School of Education, Hokkaido University
Kita 11 Nishi 7, Sapporo 060-0811 Japan

Abstract

The purpose of this study was to examine the usefulness and limitations of vector diagrams, consisting of lines with arrows representing variables, in statistical training. Nineteen undergraduates learned advanced level statistics either with vector diagrams or in the conventional way and solved three problems. Vector diagrams sometimes helped the students understand descriptions in the text which were difficult in conventional explanations, but caused other difficulties. Vector diagrams were useful for solving one of the three problems, but not the other two. It is concluded that a property of diagrams or formulae can be a double-edged sword.

Students who are majoring in psychology or other relevant disciplines have to study statistics. Despite substantial effort by teachers, understanding statistics is often difficult for many students. This paper reports the results of a practical experiment in which the students learned to employ either "vector diagrams" or a conventional formula-based approach to the basics of regression analysis. The students were then asked to solve three problems using the given technique they learned.

Unlike many previous studies on using diagrams in educational settings, which focus only on the usefulness of diagrams, this study also investigates limitations of diagrams. Research on diagrammatic reasoning has found many "good" properties of diagrams (e.g., Barwise & Etchmendy, 1996; Cheng & Simon, 1995; Larkin & Simon, 1987). The researchers seem to consider these properties as if they are always support (at least do not impair) understanding and problem solving. The results of this study suggest that the same property, which definitely makes the solution of a problem easy, sometimes makes another problem difficult. Similarly, the results suggest that formulae do not necessarily have "bad" properties.

The vector diagrams used in this study consist of several vectors drawn as lines with arrows, each of which corresponds to a variable. For example, the correlation coefficient is defined as $\cos\theta$ where θ is the angle between two vectors, $\vec{x} = \{x_1 - \bar{x}, \cdots, x_n - \bar{x}\}$ and $\vec{y} = \{y_1 - \bar{y}, \cdots, y_n - \bar{y}\}$. The regression analysis is described as the projection of the dependent variable (actually the vector of the dependent variable like $\vec{y}$ in Figure 1) on the linear space of independent variables ($\vec{x_1}$ and $\vec{x_2}$ in Figure 1).

Tasks and Prediction

These are two basic assumptions in this study about the nature of diagrammatic and algebraic representations and operators.

One assumption is that the diagrammatic representation of a problem affords a far smaller number of operators than the algebraic representations of that problem. This assumption is two-fold. First, the problem space (the set of all possible problem states) is smaller when a problem is represented by a diagram. In the chapters on vectors in mathematics textbooks, the only diagrammatic operators one can commonly find are: extension (or reduction), rotation, projection, and (de)composition. Algebraic representations, by contrast, allow many kinds of manipulations such as the four basic operations of arithmetic, expansion or factorization, fraction operations, root operations (e.g., $\sqrt{a} * \sqrt{b} = \sqrt{ab}$), summation operations (e.g., $\sum(a + b) = \sum a + \sum b$), substitution, and so on. Second, I assume that the search space (the set of problem states a student actually considers), is also smaller when a problem is represented by a diagram. Whether a problem is represented by a diagram or a formula, students do not consider all possible operators because in each case some operators are difficult to use.

The other assumption is that a formula and its transformation become more concrete when they are connected to a diagram. This assumption seems to have no room for doubt because connecting a formula to a diagram increases the number of attributes the formula has. This assumption is related to the first assumption. Because of the limited number of diagrammatic operators, "diagrammatic inference" often requires using algebraic representations, although the diagrams play a crucial role in the inference. This means that students often have to do "heterogeneous inference," inference that use multiple forms of representation (Barwise & Etchemendy, 1996).

This study claims that any property of diagrams or formulae can be either a help or a hindrance in problem solving. For example, considering only a

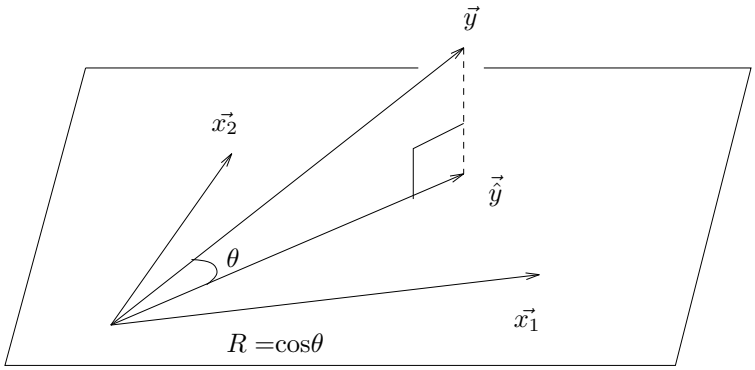

Figure 1: The definition of the multiple correlation coefficient R using a vector diagram

Table 1: Three Problems used in This Study.

Problem 1: The multiple correlation coefficient R indicates the goodness of fit of the model in multiple regression analysis. Consider the multiple correlation coefficient "R" in the case of simple regression. Please explain the relationship between this R and r, the simple correlation coefficient for the two variables x and y.
Problem 2: In the case of simple regression, if the number of paired values (x_i, y_i) is two, we can describe the values of one variable by using the other variable without any error, indicating $r = +1$ or -1. Give an explanation for the reason of this perfect description.
Problem 3: Explain the relationship between the regression coefficient $\hat{a_1}$ ($= S_{xy}/S_{xx}$) and the correlation coefficient r in the case of simple regression by using only the two variances S_{xx} and S_{yy}. S_{xy} means the covariance for the two variables x and y.

small number of diagrammatic operators can serve either as a constraint on the search (This is expected to be the case in Problem 1 in Table 1, as described below), or as a limitation if the correct solution path is outside of the search space (This is expected to be the case in Problem 2). The abstractness of formulae also can be an advantage or a disadvantage (This contrast is expected to be shown between Problem 1 and Problem 3).

Problem 1 If students learn regression analysis in a conventional way, R is defined by the formula

$$R = \frac{\frac{1}{n}\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\frac{1}{n}\sum(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum(\hat{y}_i - \bar{\hat{y}})^2}}.$$

This formula means that R is defined as the correlation coefficient between the expected value $\hat{y}$ and the observed value y. In the case of simple regression, we can say $\hat{y}_i = a_0 + a_1 x_{i1}$ and $\bar{\hat{y}} = a_0 + a_1 \bar{x}_1$. If these relations are used in the formula for R then the conclusion $R = |r|$ is obtained after a long series of algebraic manipulations.

Using a vector diagram, R is defined as $\cos\theta$ for two vectors $\vec{\hat{y}} = \{\hat{y}_1 - \bar{\hat{y}}, \cdots, \hat{y}_n - \bar{\hat{y}}\}$ and $\vec{y} = \{y_1 - \bar{y}, \cdots, y_n - \bar{y}\}$ as shown in Figure 1. Vector $\vec{\hat{y}}$ is the orthogonal projection of $\vec{y}$ on the plane spanned by $\vec{x}_1$ and $\vec{x}_2$. If we delete $\vec{x}_2$ from Figure 1 and redraw the orthogonal projection, we will obtain the answer as shown in Figure 2.

According to the basic assumption mentioned above, the problem/search space of the diagrammatic version of this problem is assumed to be smaller than the problem/search space of the formula version. The solution with vector diagram, consequently, should require less computation than the conventional solution. Note that in both solutions we started with the definition of R. In general, diagrammatic approaches often require less computation than conventional approaches (Cheng, 1992).

Other than the small problem/search space of the diagrammatic solution, the concreteness of diagrammatic operators also can contribute to finding the answer to this problem. In contrast, the formula version of the definition of R and its transformation are more abstract with no diagrammatic meaning, and the solution is a pure algebraic solution. Cheng and Simon (1995) pointed out that conventional mathematical approaches are often more complex than diagrammatic approaches because the bulk of the reasoning must center around abstract equations.

We therefore predict that a group of students which uses a vector diagram to solve this problem will show better performance than another group of students which tries to solve it in a conventional way.

In this study, after trying to solve the problems, the students read the correct solution and evaluated

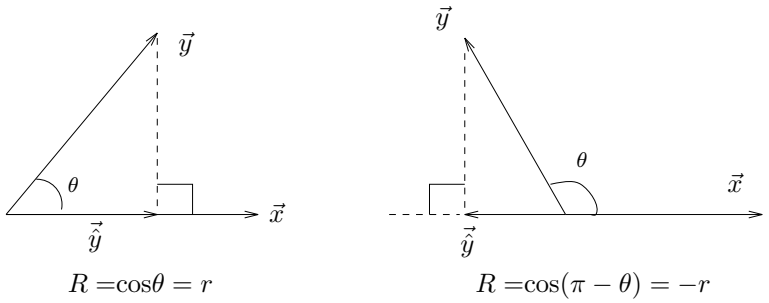

Figure 2: The answer to Problem 1 with vector diagrams

the degree of their understanding (to what degree they could understand the solution) and the degree of their conviction (to what degree they could accept it) on a 5-point scale. In Problem 1, we can expect some differences in these scores in accordance with the difference in difficulty of problem solving. However, there may be no difference in these evaluations between the two groups of the students. Even if it is difficult to make a long sequence of appropriate operators by their own efforts, just following the correct sequence may not be a tough task as long as the students are familiar with these operators.

Problem 2 This problem was chosen in this study to show that the limited number of diagrammatic operators, which is the property of vector diagrams considered to make Problem 1 easy, can also be a hindrance in problem solving. Among diagrammatic operators one can find in mathematics textbooks, I assume that the decomposition of a vector is relatively difficult to use for students because the pair of vectors that would be generated does not exist in the current problem state. If a diagrammatic solution of a problem requires students to use the decomposition operator, the correct solution path is likely to be outside of the search space, although this path is in the problem space. A crucial difference between the vector solutions of Problem 1 and 2 is in whether the correct solution path is within the search space or not, although some other differences remain uncontrolled. This experiment puts the external validity above the internal validity, and it is difficult in this kind of practical study to strictly control all factors.

In many conventional textbooks, the correlation coefficient is explained with a scatter diagram. In the case of Problem 2, two points will be plotted on the scatter diagram. The regression straight line is uniquely specified because two points define a unique line. For this problem, the comparison is not diagram vs. algebra but vector diagram vs. conventional way.

A vector diagram which can be used to solve this problem is shown in Figure 3. The vector $\vec{x}$ lies at right angles to the vector $(\bar{x}, \bar{x})$; the vector $\vec{y}$ lies at right angles to the vector $(\bar{y}, \bar{y})$. Students could find these spatial relations by drawing a diagram of concrete data chosen arbitrarily or by calculating the inner product. The fact that the vector $\vec{x}$ is parallel to the vector $\vec{y}$ means that the vector $\vec{y}$ is described as $\alpha\vec{x}$ where α is a scalar. Note that this solution needs only one kind of diagrammatic operator: Participants need to decompose each of (x_1, x_2) and (y_1, y_2) into two vectors as shown in Figure 3.

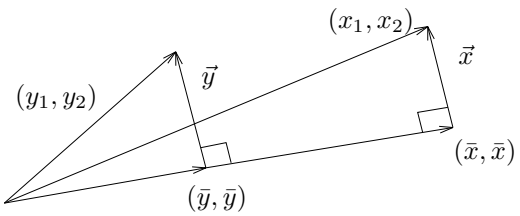

Figure 3: A vector diagram used to solve Problem 2

Our prediction is that the difficulty of using the decomposition operator will impair performance of the students. It is also expected that these students will have trouble in understanding and accepting the correct solution because the decomposition would be outside of the search space. This is contrary to the case of algebraic solution of Problem 1 because all problem states in this solution are expected to be included in the search space.

Problem 3 This problem was chosen to use in this experiment to show that the abstractness of algebraic solutions can sometimes help problem solving. Recall that it is thought that this property of formulae would make difficult the algebraic solution of Problem 1.

A conventional solution to this problem consists of a sequence of simple transformations of the equation defining the regression coefficient:

$$\hat{a}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r_{xy} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}.$$

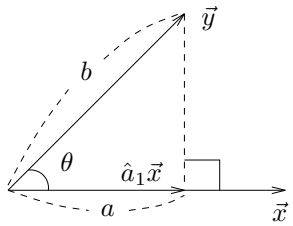

Figure 4: A vector diagram used to solve Problem 3

The transformations used in this solution look very formal and it is difficult to find any concrete meaning in them. For example, $\sqrt{S_{yy}}$ is forced to be put into the formula but this looks like a manipulation without any concrete meaning.

This problem represents a class of problems which could not be solved by purely diagrammatic thinking; rather, it requires heterogeneous inference, recruiting both diagrammatic and algebraic representations. Figure 4 shows a vector diagram which can be used to solve this problem. If two relations $a = \hat{a}_1\sqrt{\sum(x_i - \overline{x})^2}$ and $b = \sqrt{\sum(y_i - \overline{y})^2}$ are put into the equation $r = \cos\theta = \frac{a}{b}$, we will obtain the correct answer. The diagram gives some concrete meaning to this solution (the second basic assumption in this study mentioned above).

The prediction is that the students who use a purely algebraic approach will show better performance than the students who try to use a vector diagram. As mentioned in Problem 1, conventional algebraic approaches are often more complex than diagrammatic approaches because the bulk of the reasoning must center around abstract equations (Cheng & Simon, 1995). This study, however, claims that the abstractness of algebraic manipulation is not a "bad" property of formulae by nature. Heterogeneous inference requires students to use multiple representations simultaneously and it can burden students with a cognitive load. A pure diagrammatic solution, if any, is thought to be easier if this solution is as simple as the algebraic part of the heterogeneous inference.

We can expect some differences in the score of understanding and acceptance in accordance with the difference in difficulty of problem solving. However, similar to the case of Problem 1, there may be no difference in these evaluations between the two groups of students because just following the solution steps might not be very difficult.

Method

Participants

Participants were 19 undergraduate students majoring in psychology at Wakayama University, Japan. They had all taken or were taking a first introductory statistics course for psychology students, but did not know about the regression analysis taught in this experiment. In Japan, most of the undergraduate students learn algebra and vectors in high schools. This means they are ready for learning statistics either by conventional method or by an alternative, diagrammatic method.

Design

There were two experimental groups. In *formula* group, participants studied the basics of regression analysis in the conventional way in which formulae were mainly used. In *vector* group, the basics of regression analysis were taught with vector diagrams. The participants were assigned to one of these groups.

Before this experiment, participants received a simple pretest, the purpose of which was to evaluate their basic knowledge of statistics. This pretest consisted of five items, which required students to write formulae of mean, variance, SD, covariance, and correlation coefficient. Students got one point for each correct answer giving a maximum score of 5 points.

I tried to make sure that the two groups were of roughly comparable ability. Based on the results of the pretest, students were divided into nine pairs with one left over. Paired students' scores differed by a maximum of one point. Within a pair, the students were randomly assigned to one of the two groups. One remaining participant was assigned to the vector group. Thus, the vector group had 10 and the formula group had 9 participants. The two groups had roughly comparable spread of ability.

Materials

The text material was written by the author because no appropriate material was found. Two types of textual material was used corresponding to the two groups. To make these two textual materials have the same difficulty as much as possible, I first wrote the material to be used in the formula group and then "translated" it into the text used in the vector group.

Three problems shown in Table 1 were used in this experiment. All of the statistical concepts that were needed to solve these problems were explained in the textual material for both groups. Because of space limitation, I omit the detailed description of the content of these textual materials.

Procedure

There were two sessions in this experiment: understanding the text material, and problem solving. The experiment was conducted in groups of 3 to 8 participants.

In the first session, participants tried to understand the text material. If they found a description that was difficult to understand, they were required to underline that part in the textbook and to note the reason for the difficulty in the margin. The participants took about one hour to finish this session although there was no time limit.

After reading the text material, each participant reported to the experimenter (i.e. me) their difficulties in understanding the text. The experimenter gave each participant additional explanations about the difficult points in the text. All of the questions were resolved before the participants proceeded to the second session.

In the second session, the three problems shown in Table 1 were presented one by one. Fifteen minutes were allocated to solving each problem. Participants were allowed to look at the text material at any time. All participants were given a paper-and-pencil version of the test.

The participants received the correct answer after they had finished trying to solve each problem. They were required to evaluate to what degree they could understand

each solution and to what degree they could accept it on a 5-point scale ranging from 1: "very difficult to understand (or accept)" to 5: "very easy to understand (or accept)."

Results

Understanding text

It turned out that the two textual materials were similar in the sense that they had almost the same difficulty. Column 4 in Table 2 presents the number of descriptions reported as being difficult to understand in the text for each participant. There was very little difference in the number of reported difficulties during the learning session between the two groups in this experiment. The number was 8 in the formula group and 9 in the vector group.

A closer look at the reported difficulties revealed that vector diagrams often helped the students in understanding several points in which the students in the formula group had a difficulty, while other obstacles arose with vector diagrams. In the formula group, 5 of 9 participants (N, P, Q, R and S) said that it was difficult to understand the proof which showed that the range of correlation coefficient is from -1 to $+1$. Two participants (O and S) found difficulty in understanding the reason why dividing covariance by two standard deviations was the most proper way to capture the relation between two variables. One participant (R) said that she had trouble in understanding where a_0 and a_1 in the formula $\hat{y}_i = a_0 + a_1 x_i$ came from. In the vector group, all these points were not problematic for the students. No students in this group reported any difficulties in understanding the corresponding points in their textual material. Instead, they had trouble in understanding other points. Four of the 10 participants (D, G, H, and J) said that they did not know the concept of orthogonal projection (see Figure 1). Two participants (G and H) said understanding inner product—which was used in this group to define correlation coefficient—was difficult. Two participants (H and I) had difficulty in imagining n-dimensional vectors. One participant said the equation which describes the relation between the variance and a vector was difficult.

Problem solving

Columns 5, 8 and 11 in Table 2 present the performance of each participant and success S (%) in problem solving for each group; F means failure in problem solving. For each problem, the two columns to the right of the column indicating success or fail in problem solving show participants' self-evaluation of the degree of understanding and acceptance of the correct solution presented after their attempt at problem solving.

All in all, the results supported our prediction.

Problem 1 Vector representations facilitated solution of Problem 1. In the vector group, one participant (Participant F in Table 2) reached the conclusion $R = |r|$ and 5 participants found the answer $R = r$ in the case of $r \geq 0$. All of these students used vector diagrams. In the formula group, no participant got the answer $R = |r|$ or $R = r$ to this problem. The difference in success S (%) between the two groups was significant (Fisher's exact test, $p = .011$).

No significant difference was found in the self-evaluation scores for understanding and acceptance of the given correct solution.

Problem 2 and 3 In contrast to the good performance on Problem 1, no participant in the vector group succeeded in solving Problem 2 and Problem 3. The participants in the formula group showed relatively good performance. The difference in success S (%) between the two groups was significant in Fisher's exact test, $p = .003$ for Problem 2 and $p = .011$ for Problem 3.

The scores for understanding and acceptance of the correct solution to Problem 2 in the vector group were lower than the scores in the formula group. The differences in means between the two groups were significant, $t(9.0) = 5.28, p = .001$, for understanding; $t(17) = 3.15, p = .006$, for acceptance.

There was no significant difference in the scores of the two groups for understanding and acceptance in Problem 3.

Discussion

The limited number of diagrammatic operators can make the problem space smaller, and raise the probability of reaching the correct answer. We predicted that this property would improve performance on Problem 1 and the results supported this prediction. Formulae allow students to do many kinds of manipulation. For example, in Problem 1, participant R tried to get $R \times \frac{1}{r}$ and participant M considered $\frac{\{1/n \sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{R}$. Note that it is next to impossible to do these manipulations on a vector diagram. If a diagram rules out these messy manipulations, it must be a big help for students.

Interestingly, the same property, namely, affording a small number of operators, could prevent students from finding the solution and understanding an explanation. This is the case in Problem 2. No participant in the vector group succeeded in solving this problem. The participants also had trouble in understanding and accepting the correct answer to this problem. After the experiment, participant A told me that understanding the decomposition of vectors was difficult, especially, $(\bar{x}, \bar{x})$ and $(\bar{y}, \bar{y})$ looked strange. This feedback suggests that the students were likely to rule out the decomposition operator necessary

Table 2: Summary of the Data from Experiment 2.

				Problem 1			Problem 2			Problem 3		
Groups	Participants	Pre	Difficulties	S/F	Un	Ac	S/F	Un	Ac	S/F	Un	Ac
Vector	A	3	0	F	5	2	F	2	1	F	4	4
	B	3	1	S	4	4	F	2	3	F	2	2
	C	3	0	S	5	5	F	4	3	F	5	5
	D	1	1	F	1	1	F	4	2	F	4	2
	E	1	0	S	5	5	F	2	3	F	5	5
	F	1	0	S	4	4	F	4	4	F	4	4
	G	1	2	F	3	3	F	1	1	F	5	5
	H	1	3	F	4	3	F	1	1	F	1	1
	I	1	1	S	4	4	F	4	4	F	4	4
	J	1	1	S	4	5	F	4	4	F	4	4
	Mean/%correct	1.60	0.90	60.0%	3.90	3.60	0.0%	2.80	2.60	0.0%	3.80	3.60
	SD	0.92	0.94		1.14	1.28		1.25	1.20		1.25	1.36
Formula	K	3	0	F	4	3	F	5	3	F	5	4
	L	3	0	F	2	4	S	5	5	S	5	5
	M	2	0	F	4	3	S	5	5	S	4	3
	N	2	1	F	4	4	S	5	5	F	5	5
	O	1	1	F	5	5	F	5	4	S	5	5
	P	1	1	F	4	4	S	5	5	F	4	2
	Q	1	1	F	5	5	F	5	2	S	3	3
	R	1	2	F	4	4	S	5	5	S	4	4
	S	0	2	F	4	2	S	5	5	F	4	4
	Mean/%correct	1.56	0.89	0.0%	4.00	3.78	66.7%	5.00	4.33	55.6%	4.33	3.89
	SD	0.96	0.74		0.82	0.92		0.00	1.05		0.67	0.99

Notes. Pre: the score of pretest (1–5)
Difficulties: the number of descriptions in the text which were difficult to understand
S/F: success (S) or failure (F) in problem solving
Un: the score of evaluating the degree of understanding the correct solution (1–5)
Ac: the score of evaluating the degree of acceptance of the correct solution (1–5)

to solve this problem. The low ratings for understanding and acceptance of the correct answer refute the possibility that the inability to solve this problem means that the participants carelessly failed to apply a familiar operator.

Similar to the case of properties of diagrams, a property of formulae can be either an advantage or a disadvantage. Abstractness is an example of such properties. This property was predicted to work against solving Problem 1 but to be an aid in solving Problem 3 in the formula group. The results of the experiment were consistent with this prediction. A formula and its transformation become more concrete when they are connected to a diagram. This is the case in heterogeneous inference, inference that use both diagrammatic and algebraic representations. A pure diagrammatic solution is easier if this solution is as simple as the algebraic part of the heterogeneous inference.

Conclusion

Previous research on diagrammatic reasoning has pointed out many "good" properties of diagrams and has claimed advantage for diagrammatic approaches over conventional (usually algebraic) approaches. From the results presented here, it seems that the story is not so simple. The results of this experiment indicate that the vector diagram is not a panacea for students struggling with statistics. The same property of certain diagrams or formulae can be either an advantage or a disadvantage. Teachers should keep this in mind and ponder how properties of diagrams or formulae can work in a particular situation.

Acknowledgement

I wish to thank Sciencedit (http://www.sciencedit.com/) and two of my friends, Ryan Baker and Erik Lindsley, for editing the manuscript of this paper.

References

Barwise, J., & Etchmendy, J. (1996). Visual information and valid reasoning. In G. Allwein & J. Barwise (Eds.), *Logical Reasoning with Diagrams*. New York: Oxford University Press.

Cheng, P. C. -H. (1992). Diagrammatic reasoning in scientific discovery: Modeling Galileo's kinematic diagrams. *Reasoning with Diagrammatic Representations: Papers from the 1992 Spring Symposium*. (pp. 33–38) Menlo Park, CA: AAAI Press

Cheng, P. C. -H., & Simon, H.A. (1995). Scientific Discovery and Creative Reasoning with Diagrams. In S. M. Smith, T. B. Ward & R. A. Finke (Eds.), *The Creative Cognition Approach*. Cambridge, MA: MIT Press.

Larkin, J., & Simon, H.A (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99.