Counterexample search in diagram-based geometric reasoning

Yacin Hamami

Centre for Logic and Philosophy of Science, Vrije Universiteit Brussel

John Mumma

Philosophy Department, California State University of San Bernardino

Marie Amalric

CAOs Laboratory, Department of Psychology, Carnegie Mellon University

Final Version Submitted to *Cognitive Science*

Author Note

Corresponding authors:

Yacin Hamami, Pleinlaan 2, B-1050 Brussels, Belgium,

Email: yacin.hamami@gmail.com.

Marie Amalric, 5000 Forbes Ave, Pittsburgh, PA 15213, USA,

Email: marie.amalric@normalesup.org.

Abstract

Topological relations such as inside, outside, or intersection are ubiquitous to our spatial thinking. Here, we examined how people reason deductively with topological relations between points, lines, and circles in geometric diagrams. We hypothesized in particular that a counterexample search generally underlies this type of reasoning. We first verified that educated adults without specific math training were able to produce correct diagrammatic representations contained in the premisses of an inference. Our first experiment then revealed that subjects who correctly judged an inference as invalid almost always produced a counterexample to support their answer. Noticeably, even if the counterexample always bore a certain level of similarity to the initial diagram, we observed that an object was more likely to be varied between the two drawings if it was present in the conclusion of the inference. Experiments 2 and 3 then directly probed counterexample search. While participants were asked to evaluate a conclusion on the basis of a given diagram and some premisses, we modulated the difficulty of reaching a counterexample from the diagram. Our results indicate that both decreasing the counterexample density and increasing the counterexample distance impaired reasoning performance. Taken together, our results suggest that a search procedure for counterexamples, which proceeds object-wise, could underlie diagram-based geometric reasoning. Transposing points, lines, and circles to our spatial environment, the present study may ultimately provide insights on how humans reason about topological relations between positions, paths, and regions.

Keywords: Diagram-based geometric reasoning ; Counterexample search ; Topological relations ; Geometric cognition ; Mathematical reasoning ; Spatial reasoning.

Counterexample search in diagram-based geometric reasoning

1 Introduction

Thinking and reasoning about entities in the space surrounding us is one of our essential cognitive abilities and a topic of intense research in cognitive sciences (see, e.g., Burgess, 2008; Hegarty & Stull, 2012; Majid, Bowerman, Kita, Haun, & Levinson, 2004; Newcombe, 2018; Shah & Miyake, 2005; Tversky, 2005; Wang & Spelke, 2002). Among our spatial cognitive abilities, our capacity for spatial deductive reasoning allows us to infer new information about the spatial arrangements of objects from information we may have obtained through observation or communication. For instance, being told that "the toolbox is behind the car" and having observed that "the car is behind the truck", you may safely infer that "the toolbox is behind the truck". Spatial deductive reasoning of this kind has been extensively studied in the psychology of reasoning (see, e.g., De Soto, London, & Handel, 1965; Byrne & Johnson-Laird, 1989; Van der Henst, 2002; Knauff, 2013; Ragni & Knauff, 2013; for a review of earlier work see Evans, Newstead, & Byrne, 1993, chapter 6). However, previous studies have almost exclusively focused on discrete positional relations (e.g., to the left of, to the right of, above, below, behind, in front of, etc.) or on one-dimensional topological relations (Knauff, 1999; Knauff, Strube, Jola, Rauh, & Schlieder, 2004), but have not typically studied topological relations involving regions and their boundaries (such as being inside, outside, or intersecting) that are yet fundamental to our spatial understanding.

By its very nature, diagram-based geometric reasoning constitutes both a privileged and familiar setting to study reasoning with topological information since geometric objects such as points, lines, and circles are among the simplest entities to support inferences involving topological relations. Reasoning with geometric diagrams is also one of the oldest form of mathematical and scientific thinking going back to the revolutionary work of Thales, Euclid, and Archimedes in ancient Greece (Netz, 1999). It played an important role in various human intellectual achievements in astronomy, architecture, engineering, and land management (Kline, 1972). It remains today the way that many students are introduced to the method of deductive proof in mathematics (NCTM, 2000). But while it has been extensively studied by historians (Mueller, 1981; Netz, 1999), philosophers (Giaquinto, 2011; Macbeth, 2010; Manders, 2008;

Panza, 2012), and logicians (Avigad, Dean, & Mumma, 2009; Miller, 2007; Mumma, 2006), it has received only very little attention in experimental psychology so far (for a notable exception, see Koedinger, 1991; Koedinger & Anderson, 1990).

In the past decades, cognitive studies conducted in children and adults of various cultural backgrounds have revealed that all humans possess capacities of intuitive geometric reasoning. For example, Amalric et al. (2017) showed that all humans, regardless of their age and level of education, spontaneously use rotations and axial symmetries to detect and predict regularities in geometric spatial sequences. Moreover, Amazonian Mundurukú people, even if largely deprived of formal schooling and of numerical and geometric lexicon, nonetheless exhibited intuitions about the intersection of two lines extended indefinitely, or about the alignment of some points in the plane and to a lesser extent on a sphere (Izard, Pica, Spelke, & Dehaene, 2011). They could also successfully complete a triangle from its base with the appropriate angle (Izard et al., 2011; see also Hart et al., 2018), and proved able to read and use geometric information contained in abstract maps in order to locate a target object, even though it was the first time they were presented with such tools (Dehaene, Izard, Pica, & Spelke, 2006; see also Dillon, Huang, & Spelke, 2013; Izard, O'Donnell, & Spelke, 2014).

At a radically different level, some cognitive studies attempted to describe and model expert geometric reasoning. Koedinger and Anderson (1990) have monitored the successive steps that experts in geometry take to solve classical problems of triangle geometry (see also Kao, Douglass, Fincham, & Anderson, 2008; Koedinger, 1991). They notably found that experts consistently focus on the same important steps and skip the same minor ones. Interestingly, these important steps seem to correspond to perceptual chunks on the geometric diagrams corresponding to the problems. These observations led to the construction of a model of expertise in geometric proof problem solving that has been used to inform the development of cognitive tutors (Koedinger & Anderson, 1993; Ritter, Anderson, Koedinger, & Corbett, 2007). Pursuing in this direction, Koedinger (1998) has characterized the cognitive skills in conjecturing and proving to be acquired in a high-school geometry course, and has made proposals on how to use interactive geometry software to enhance the learning of these skills.

Diagram-based geometric reasoning—understood here as reasoning with topological relations between geometric objects through geometric diagrams (see Figure 1)—lies between these different forms of geometric reasoning. On the one hand, it is more specific than the expert geometric reasoning just discussed in that it concerns individual inferences and not the process of finding a sequence of inferences for solving a geometric problem. On the other hand, it is more general than the different kinds of intuitive geometric reasoning previously studied in that it does not concern particular or specific geometric objects presented visually but rather classes of geometric objects standing in certain relations which are stated in linguistic form. Yet it still appears as a form of intuitive geometric reasoning insofar as it does not require advanced geometric expertise of the kind required to find geometric proofs.

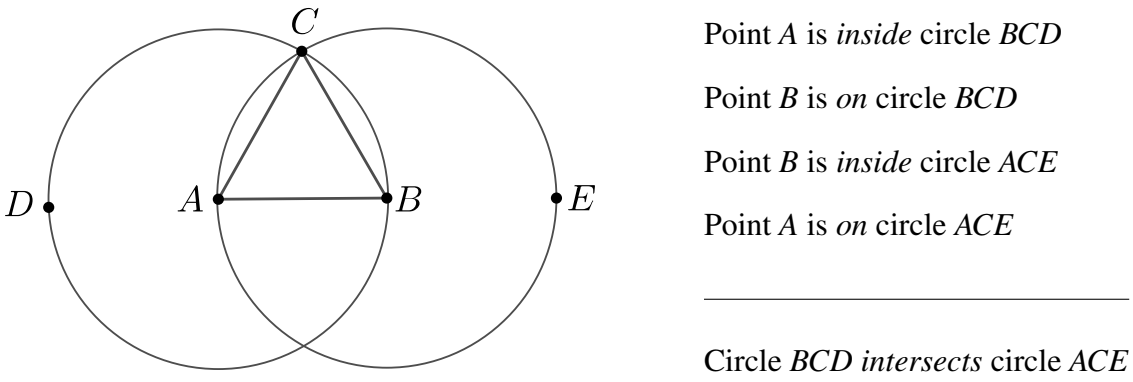


Figure 1. Diagram accompanying the proof of Proposition 1 from Book I of Euclid's *Elements*. This proposition shows how to construct an equilateral triangle on a given finite straight line with only a compass and a ruler (Euclid, 1959, pp. 241-242). To go through, the proof requires to infer from the diagram that the two constructed circles intersect. This is a typical diagram-based geometric inference where premisses and conclusion consist of topological relations between geometric objects.

The cognitive roles and functions of diagrams have been investigated with respect to many forms of reasoning such as syllogistic reasoning (e.g., Stenning & Oberlander, 1995; Stenning & Yule, 1997), mechanical reasoning (e.g., Hegarty, 1992, 2004; Heiser & Tversky, 2006), double-disjunctive reasoning (e.g., Bauer & Johnson-Laird, 1993), and analogical reasoning (e.g., Pedone, Hummel, & Holyoak, 2001), among others (for reviews, see Hegarty & Stull,

2012; Shah & Miyake, 2005; Tversky, 2005). Landy and Goldstone (2007a, 2007b) have argued that symbolic reasoning in algebra and logic is, in some respects, akin to diagrammatic reasoning since it relies on spatial properties of formal notations. In an educational context, Koedinger and Terao (2002) and Booth and Koedinger (2012) have also investigated the potential advantages and drawbacks of using diagrammatic representations in solving elementary algebra problems. However, the specific use of diagrammatic reasoning in elementary Euclidean geometry has yet to be addressed by cognitive sciences. Thanks to recent advances in philosophy (Manders, 2008) and logic (Avigad et al., 2009; Miller, 2007; Mumma, 2006) that provide a precise characterization of diagram-based geometric reasoning, it is now amenable to experimental investigation.

In this study, we investigated the capacity of educated adults to reason with topological relations between geometric objects through geometric diagrams. A skill central to mathematical and scientific thinking, and essential in common sense reasoning and critical thinking, is the ability to find counterexamples—this is the main method to show that a deductive argument is invalid. This ability is an important target for mathematics educators (see, e.g., Weber, 2009; Zaslavsky & Ron, 1998; Zazkis & Chernoff, 2008), and is intimately connected with rational thinking, for instance in the context of Wason's (1968) selection task. The ability to find counterexamples plays a central role in some psychological theories of human reasoning such as the mental model theory (Johnson-Laird, 2006, 2010), but only few studies have addressed it directly (see, e.g., Bucciarelli & Johnson-Laird, 1999; Johnson-Laird & Hasson, 2003), and it has been left out of some other theories based on formal rules (Braine & O'Brien, 1998; Rips, 1994), probabilities (Oaksford & Chater, 2001, 2007), or verbal reasoning (Polk & Newell, 1995). In fact, as R. M. J. Byrne, Espino, and Santamaria (1999) pointed out: "it has proved more difficult to examine experimentally the search for counterexamples, and it remains the case that little is known about how counterexample search is carried out" (R. M. J. Byrne et al., 1999, p. 348). Here, we tested the hypothesis that counterexample search underlies diagram-based geometric reasoning. Alternative hypotheses would be for diagram-based geometric reasoning to rely on mental rules (Braine & O'Brien, 1998; Rips, 1994), probabilities (Oaksford & Chater, 2001, 2007), or verbal reasoning (Polk & Newell, 1995), in which case the result-

ing cognitive accounts would not appeal to any form of search procedure for counterexamples. Among contemporary psychological theories of human deductive reasoning, only the mental model theory postulates a role for counterexample search within the reasoning process proper. However, we shall see that the counterexample search procedure we identify presents some differences with that postulated by the mental model theory. The main objectives of this study were thus (1) to investigate whether people evaluate the validity of diagram-based geometric inferences by searching for counterexamples, and if so (2) to examine the underlying search procedure. To this end, we conducted three experiments in which educated adults engaged in reasoning with geometric diagrams that they had constructed (Experiment 1) or that were provided to them (Experiments 2 and 3).

2 Formal Characterization of Diagram-Based Geometric Inferences

The formal characterization of diagram-based geometric inferences adopted in this study is based on the formal system developed by Avigad et al. (2009). Following this system, we considered a formal language $\mathscr{L}$ with three types of objects:

- *points* denoted by A, B, C, etc.,
- *lines* denoted by L, M, N, etc.,
- *circles* denoted by α, β, γ, etc.,

together with the following set of relations:

- point A is $\{$inside, on, outside$\}$ circle α,
- point A is $\{$on, off$\}$ line L,
- points A and B are $\{$on the same side, on opposite sides$\}$ of line L,
- point B is $\{$between, not between$\}$ points A and C on line L,
- line L $\{$intersects, does not intersect$\}$ line M,
- line L $\{$intersects, does not intersect$\}$ circle α,

- circle α {intersects, does not intersect} circle β,

- circle α is {inside, outside} circle β.

A proposition in the language $\mathscr{L}$ is always an atomic formula consisting of a single relation between particular geometric objects, e.g., "point *A* is inside circle α" or "circle α intersects circle β".

In this setting a *diagram-based geometric inference* **I** is entirely characterized by a set of premisses and a conclusion in the language $\mathscr{L}$. Here are two examples of such inferences:

(I_1)

Point *A* is *inside* circle α

Points *A* is *on* line *L*

Line *L intersects* circle α

(I_2)

Point *A* is *outside* circle α

Points *A* is *on* line *L*

Line *L intersects* circle α

In order to say when a diagram-based geometric inference is valid or invalid, we have to define the notion of a model (in the logical sense) for a set of propositions in $\mathscr{L}$. A *model* for a set of propositions Φ in $\mathscr{L}$ is a geometric configuration involving only the geometric objects mentioned in Φ and in which all the propositions in Φ are true. A *geometric configuration* is here defined as a set of geometric objects in the Euclidean plane. We then say that a diagram-based geometric inference **I** is *valid* whenever its conclusion is true in all the models of its premisses, and *invalid* otherwise. A geometric configuration in which the premisses of **I** are true and the conclusion is false is called a *counterexample* to **I**. Thus, a diagram-based geometric inference is valid if there are no counterexamples to it, and invalid when there exists such a counterexample. In the two examples above, the inference I_1 is valid while the inference I_2 is invalid. Figure 2 illustrates the notions of model and counterexample in the cases of inferences I_1 and I_2.

In the experiments reported below, we considered different sets of diagram-based geometric inferences. Each set always contained an equal number of valid and invalid inferences, and parameters such as the number of objects and the number of premisses were systematically varied.

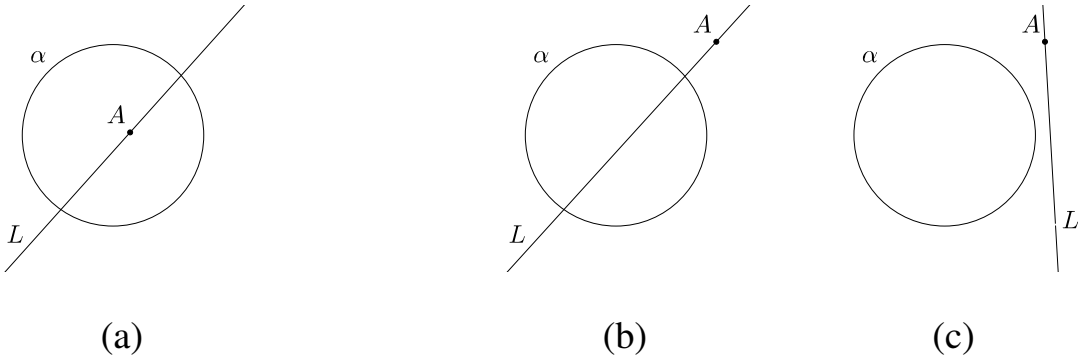


Figure 2. Three geometric configurations involving a point A, a line L, and a circle α. (a) A model of the premisses and the conclusion of I_1. (b) A model of the premisses and the conclusion of I_2. (c) A counterexample to I_2.

3 Experiment 1

In this paper-and-pencil experiment, we first verified that educated adults were capable of evaluating the validity of diagram-based geometric inferences. We then investigated whether their reasoning relies on a search for counterexamples. If this is the case, we expected them (1) to be able to provide a counterexample to an inference they judged as invalid, and (2) to judge an inference as valid when they failed to find a counterexample to it. Failure to find a counterexample to an invalid inference would lead to mistakenly classify it as valid. Such a mistake is not possible in the case of a valid inference. For this reason, we expected adults to judge more accurately valid inferences as compared to invalid ones.

We thus confronted participants with a set of geometric reasoning problems. For each problem, we first evaluated their ability to draw a possible diagrammatic representation of a situation described by a set of premisses. We then asked them to determine which relation held among certain objects of their drawing and to judge whether this relation necessarily followed from the premisses. When they answered negatively—i.e., when they judged the inference as invalid—they were asked to illustrate their answer with a second drawing. This last question allowed us to evaluate the participants' ability to produce counterexamples to invalid inferences, without directly asking for one. Finally, the systematic comparison of the first and second drawings allowed us to identify some characteristics of the counterexample production.

3.1 Method

3.1.1 Design and Materials. Participants carried out 18 reasoning problems (see Table S1, supplementary materials), half valid, and the other half invalid. The problems concerned either 3, 4, or 5 geometric objects, equally distributed. All the problems with 3 objects had 2 premisses; the problems with 4 and 5 objects had either 2, 3, or 4 premisses, equally distributed. The order of presentation of the problems was randomized.

The 18 reasoning problems were presented in a booklet, each of them on a double page of paper. On the left page, the objects and the premisses of the problems were listed, e.g., "We consider a situation concerning a point A, a line L, a circle α such that: point A is *inside* circle α; point A is *on* line L" (inference I_1). The participants were then asked to draw a possible representation of that situation. We noted in the instructions that the lines in the problems could always be extended indefinitely at each of their extremities. While the participants were working on the left page, the right page was masked by a white sheet of paper. On the right page, once uncovered, the participants were asked to choose which relation held among some objects of their first drawing, e.g., "Which of these options correspond to your drawing: line L *intersects* circle α; line L *does not intersect* circle α". Then the participants had to decide whether this relation was necessarily the case in the situation described by the premisses. Finally, if they answered negatively to this reasoning question, they were asked to illustrate their answer with a (second) drawing.

3.1.2 Procedure. The participants were tested individually in a quiet room. Each participant received a booklet, a pencil, a ruler, and a compass. They were instructed to use the ruler and compass provided to produce the drawings, and to carefully label each object in their drawings. The participants were asked to answer systematically, and when uncertain, to opt for the most plausible answer. They were also told to maintain the white sheet of paper masking the second page of each problem until they were done working on the first page. They could work at their own pace, and were specifically told never to go back to a previous problem. They were not allowed to write or draw on additional pieces of paper. The experiment lasted between 20 and 40 minutes depending on each participant.

3.1.3 Participants. The 24 participants who took part in this experiment were French speaking adults (15 female, age range: 18-75, mean: 33.4) from the Paris urban area. They had very diverse university backgrounds such as biology, philosophy, economy, physics, literature, psychology, cognitive sciences, geology, human resources, geopolitics, cinema, etc. But none of them were studying or had studied math *per se* at the university. They were recruited via the RISC platform ("Relais d'Information sur les Sciences Cognitives") and they were compensated 5 euros for their participation. The experiment was performed according to the Declaration of Helsinki (2013) such that all participants provided informed consent prior to the experiment.

3.1.4 Statistical Analysis. We used the software environment R (R Core Team, 2018) and the software RStudio (RStudio Team, 2016) for descriptive and quantitative statistical analyses. In particular, mixed effect regressions were performed with the R package lme4, and post-hoc tests with the package emmeans.

3.2 Results

We first assessed participants' ability to make correct diagrammatic representations of the geometric situations by evaluating for each problem whether the premisses were correctly represented. To this end, each drawing was attributed a score corresponding to the proportion of premisses correctly instantiated in the diagram. Across all participants and problems, 92.3 $\pm$ 4.2% of the first drawings were correct. We performed a logistic regression on the drawing scores with the problem validity, the numbers of objects and the number of premisses as fixed effects and the subjects as random effect. No effect of validity was found (valid problems: 91.0 $\pm$ 4.6%; invalid problems: 93.6 $\pm$ 3.7%; $t(23) = 1.63$, n.s.; see Figure 3). The number of premisses did not affect the correctness of the drawings either (2 premisses: 93.3 $\pm$ 4.5%; 3 premisses: 89.6 $\pm$ 3.8%; 4 premisses: 92.4 $\pm$ 3.6%; $ts < 1.1$, n.s.). The number of objects did not have any impact either on the ability of participants to draw a correct representation of the premisses (3 objects: 95.5 $\pm$ 3.5%; 4 objects: 89.1 $\pm$ 5.2%; 5 objects: 92.3 $\pm$ 3.5%); $ts < 1.4$, n.s.).

We then evaluated whether each geometric situation described in our problems tended to

be represented in a common manner by the participants. By design, there is only one possible topological configuration compatible with the premisses for valid inferences, while there are multiple ones for invalid inferences. Supplementary Table S2 shows that participants tended to prefer one of the topological configurations for the invalid problems 4, 6, 12, 14, and 18.

We then analyzed participants' performance in determining whether the conclusion of a problem necessarily followed from the premisses. To do so, only the problems that participants perfectly understood could be taken into account. We thus restricted our analysis to the cases for which the first drawing was a correct representation of the premisses (hence discarding 7.7% of the data). In these cases, participants answered the reasoning question 87.8 $\pm$ 6.8% correctly. Again, we applied a logistic regression on the reasoning scores (0 or 1) with the problem validity, the number of objects and the number of premisses as fixed effects and the subjects as random effect. This time, we found a significant effect of validity on participants' accuracy in response to the reasoning question (valid problems: 96.0 $\pm$ 4.1%; invalid problems: 80.1 $\pm$ 8.3%; $t(23) = 3.63$, $p < 0.005$). Problems with 4 premisses also appeared to induce more errors (% correct for 2 premisses: 89.4 $\pm$ 6.4%; 3 premisses: 94.2 $\pm$ 4.9%; 4 premisses: 77.9 $\pm$ 8.7%; $t(20) = 2.79$, $p < 0.05$ when comparing 3 and 4 premisses). Finally, problems with 4 and 5 objects did not induce more errors in the reasoning question than problems with 3 objects (3 objects: 88.7 $\pm$ 6.6%; 4 objects: 84.2 $\pm$ 7.6%; 5 objects: 90.4 $\pm$ 6.2%; $ts < 1.6$, n.s.).

In the cases where a problem was correctly judged as invalid, the participants almost always provided a counterexample to it. We evaluated the correctness of the second drawing following the same procedure as for the first drawing. The percentage of correct second drawings given a correct first drawing and a correct response to the reasoning question was equal to 96.6 $\pm$ 3.8% (see Figure 3). More specifically, for all invalid problems but one, participants who correctly answered the reasoning question justified their answer by providing a correct counterexample in at least 94.7% of the cases (see Table S2, sup. mat.).

When possible—i.e., in the case of problems correctly classified as invalid after a correct first drawing—we evaluated the extent to which the second drawing diverged from the first drawing. Note that this evaluation cannot be done in the case of valid problems that do not present any counterexample. To do so quantitatively, we superposed the two drawings with re-

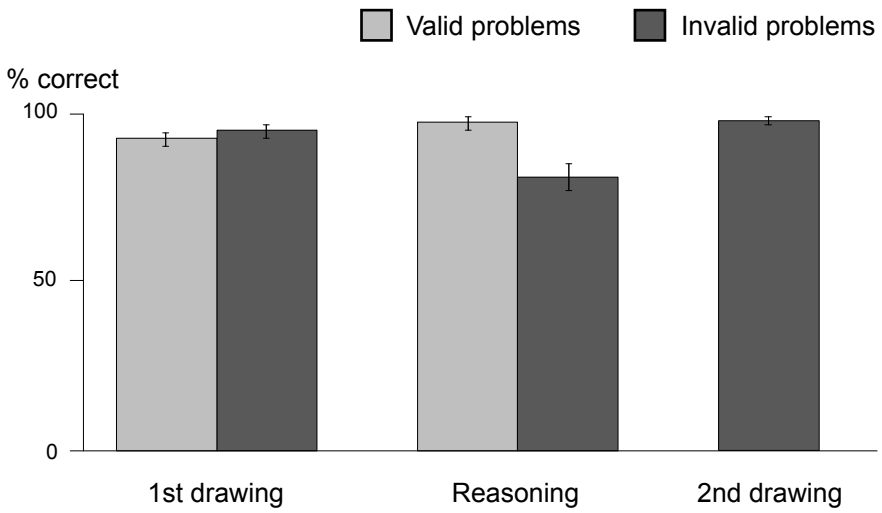


Figure 3. Percentages of correct first drawings, correct responses to the reasoning question given a correct first drawing, and correct second drawings given a correct response to the reasoning question and a correct first drawing, for the reasoning problems of Experiment 1.

spect to the maximum number of objects. The superposed objects constituted a fixed reference frame (i.e., 0 transformation), with respect to which we evaluated the number of transformations applied to each remaining object (translation, rotation, uniform scaling). Finally, the variation rate of each object in each problem was computed as the total number of transformations relative to the number of participants. The overall variation rate was equal to 46.5%. A significant difference was found between the variation rates of objects present only in the premisses and objects present in the conclusions (premisses: 29.0%; conclusion: 56.3%; $t(34) = 2.59$; $p = 0.014$; see Figure 4). For each participant, we also evaluated whether their propensity to make changes between the first and second drawings could reflect their performance in the reasoning task. We thus computed the overall percentage of correct responses in the reasoning task and the overall object variation rate for each participant, and found a small correlation between these two variables ($R(21) = 0.41$, $p = 0.054$).

3.3 Discussion

In this experiment, most of our participants were able to: (1) produce, with a high rate of success, a correct possible representation of a geometric situation described by a set of premisses; (2) discriminate between valid and invalid diagram-based geometric inferences;

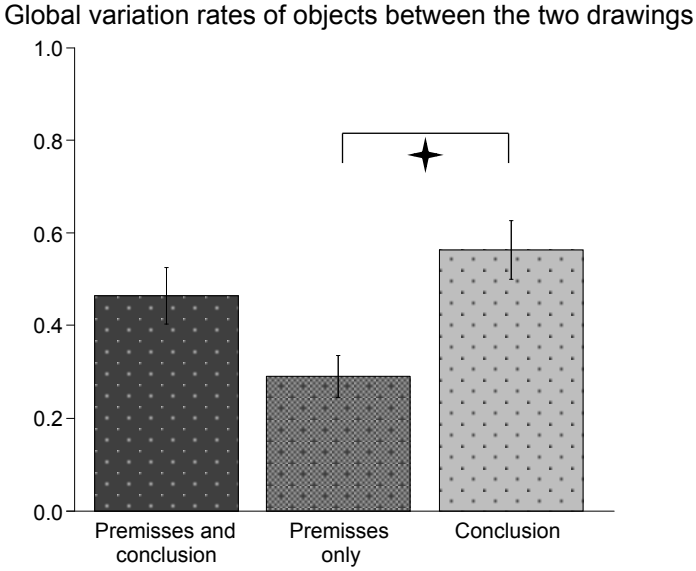


Figure 4. Global variation rates between the first and second drawings for (a) all objects, (b) objects only present in the premisses, (c) objects present in the conclusion, for the invalid problems of Experiment 1. Only cases with a correct first drawing, a correct response to the reasoning question, and a correct second drawing are taken into account.

(3) produce counterexamples to invalid inferences.

The results of this experiment support the hypothesis that educated adults evaluate the validity of diagram-based geometric inferences by searching for counterexamples. First, we saw that participants who judged an inference as invalid after having successfully represented the premisses almost always justified their answer by providing a counterexample. Second, when participants provided a correct first drawing, we observed a very high level of accuracy in classifying the problems as valid, and a significantly lower level for the invalid ones.

Our results also provide some information on how the search for counterexamples proceeds. By analyzing the differences between the first and second drawings in the case of invalid inferences, we observed a certain level of similarity between them, the likelihood for an object to remain identical between the two drawings being higher than 50%. This may indicate that the search for counterexamples proceeds locally at the level of certain selected objects. We also observed that an object was more likely to be varied between the two drawings if it was present in the conclusion as compared to being present only in the premisses. This suggests that an object is more likely to be selected in the local search if it is present in the conclusion.

This may not be surprising since a counterexample is, by definition, a geometric configuration in which the conclusion is false. Thus, for a local search to succeed, it must concern at least one of the objects present in the conclusion.

On the basis of these results, we hypothesize that the search for counterexamples proceeds by varying mentally certain object(s) in the diagram, according to the premisses they are subject to, and with the objective of falsifying the conclusion of the considered inference. In the following, we call these local procedures *scanning operations*. We note here that such scanning operations are different from the image scanning studied by Kosslyn, Ball, and Reiser (1978) that concerns the scanning of static visual images, while the scanning operations we hypothesize concern the scanning of possibilities to place one or several geometric objects in a geometric diagram, thus going beyond the information contained in the diagram. The scanning operations have a similar character to the *augmented diagram inferences* of Mumma (2012) in which the objects varied in diagrammatic inferences are those added by geometric constructions (on geometric constructions, see also Matsuda & VanLehn, 2004).

4 Experiment 2

The results of Experiment 1 led us to propose that searching for counterexamples to diagram-based geometric inferences proceeds through scanning operations. As an illustration of this procedure, suppose that one is searching for a counterexample to inference I_2 on the basis of the diagram displayed in Figure 2(b). One could proceed by carrying out a scanning operation with line L or with circle α. In the first case, the scanning operation would consist in mentally varying line L, according to the constraint that point A remains on L, and with the objective of placing L so that it does *not* intersect circle α. This scanning operation is illustrated in Figure 5(a). In the second case, the scanning operation would consist in varying circle α, with the constraint that point A remains outside α, and with the objective of placing α so that it does *not* intersect line L. This scanning operation is illustrated in Figure 5(b). In both cases, these scanning operations yield a counterexample to the inference I_2 by varying only one of the objects involved in the inference while the others remain fixed.

If adults evaluate the validity of diagram-based geometric inferences by searching for

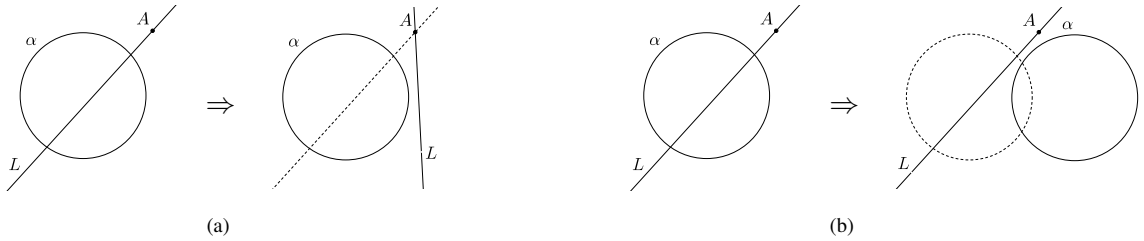


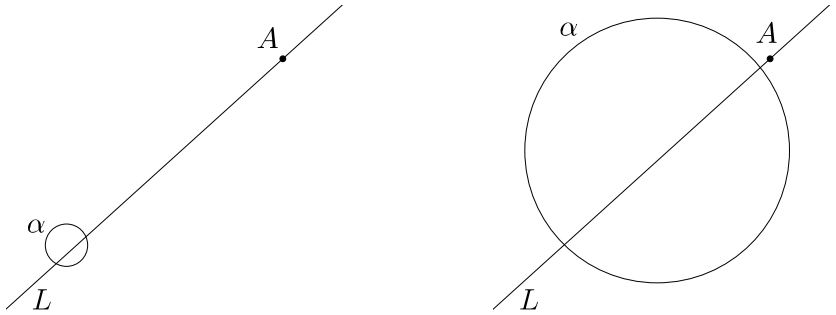
Figure 5. Two examples of scanning operations starting with the diagram displayed in Figure 2(b). (a) Scanning operation with line *L*. (b) Scanning operation with circle α.

counterexamples, then we would expect that manipulating the difficulty of finding counterexamples in the diagram through scanning operations would affect their performance in detecting invalid inferences. To test this prediction, we introduced a set of *scanning problems* that consisted in answering a reasoning question about an object introduced on a pre-existing diagram and subject to one or more constraints. We particularly investigated the effect on reasoning performance of two different metric manipulations of the provided diagram: variation of the counterexample density, and variation of the counterexample distance. The former is the subject of this Experiment 2, while the latter is the subject of Experiment 3.

We define *counterexample density* as the proportion of counterexamples obtained when considering all possible ways of placing the object to which the reasoning question applies in the diagram according to the constraints. The counterexample density can thus be viewed as the probability to "hit" a counterexample by placing randomly the object in the diagram according to the constraints. Figure 6 illustrates the notions of counterexample density in the case of inference I_2 when considering a scanning problem with line *L*. When inferences are accompanied by diagrams of high counterexample density, we expect reasoning to be facilitated.

4.1 Method

4.1.1 Design and Materials. In this experiment, we presented 12 invalid scanning problems, that were seen once with a diagram of low counterexample density, and once with a diagram of high counterexample density. To counterbalance the answer to the reasoning question, we also presented 12 valid problems, seen in two different diagram conditions (see Tables S3 and S4, sup. mat.). These 24 scanning problems were selected based on the 18 reasoning



(a) (b)

Figure 6. Two possible diagrams in the case of inference I_2. When considering a scanning problem with line *L*, (a) is a diagram of *high* counterexample density and (b) is a diagram of *low* counterexample density.

problems presented in our first experiment and 6 additional problems. They involved either 3, 4, or 5 geometric objects, equally distributed.

The 48 problems were presented to each participant in a pseudo-random order such that: (1) the order in which the two diagram conditions were seen for each problem was properly counterbalanced, so as to overcome an eventual effect of learning during the test; (2) it was very unlikely for the two diagram conditions of the same problem to be presented close to each other; (3) all the problems were presented in a different random order to each participant.

On each trial, a partial diagram was displayed at the center of the screen with ample margins around it, in a block of 300x300 pixels within a larger block of 900x500 pixels. After 3 seconds, an additional object appeared in the diagram, together with one or more premisses expressing the constraint(s) it is subject to. The participants were given 2.5 seconds plus 500 ms per word to read the premisses and look at the diagram (see Figure 7). They were finally prompted to answer "Yes" or "No" to a question about the object that last appeared. The reasoning time was measured as the delay between the appearance of the question and participants' click on the chosen response box. If they failed to answer within 10 seconds, the test automatically switched to the next problem.

To familiarize the participants with the task, the experiment was preceded by two training series, involving a separate set of 4 problems similar to those of the main test, presented without time limits.

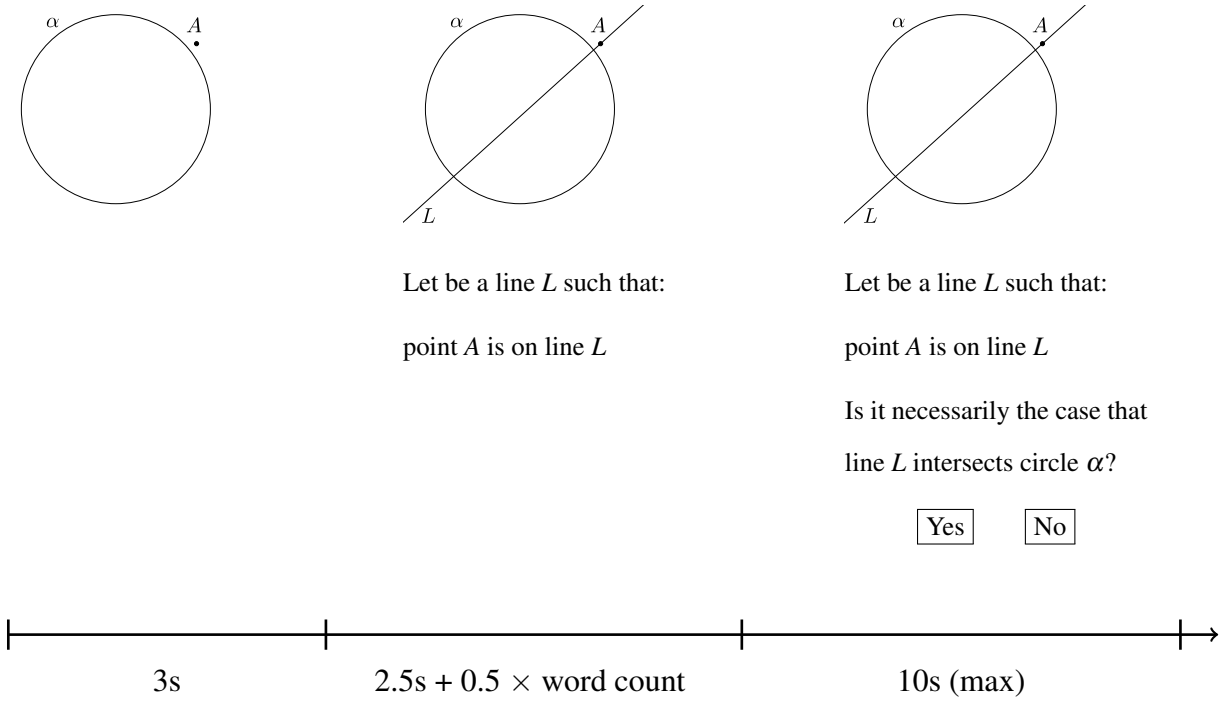


Figure 7. The three presentation stages of each scanning problem: (1) display of a first partial diagram; (2) addition of an object accompanied by a text describing its constraints; (3) display of the reasoning question and response period.

The experiment was programmed in HTML, CSS, and JavaScript using the jsPsych library (de Leeuw, 2015). It was run online on the Amazon Mechanical Turk platform. We used JATOS (Lange, Kühn, & Filevich, 2015) on the server side to manage participants and collect data.

4.1.2 Procedure. The participants were tested online. They were told to place themselves in a quiet environment because the test would require their full attention. They were also asked not to do anything else in their web browser in parallel with the test. The experiment was administered in French. To participate, one had to successfully pass a French language test composed of 8 reading and comprehension questions. The participants were then instructed that they would have to answer questions about various geometric situations. They were asked to provide their responses as quickly as possible, and were told that each trial would time out after 10 seconds.

4.1.3 Participants. 26 French speaking adults (8 females, age range: 20-53, mean: 35.6) were recruited on the Amazon Mechanical Turk platform with the two followings worker qualifications: “HIT Approval Rate (%) for all Requesters’ HITs greater than 95” and “Location

is France" or "Location is Canada". If they carried out the test in its entirety, they received $2 for their participation. According to the Declaration of Helsinki (2013), all participants provided informed consent prior to the experiment.

4.1.4 Statistical Analysis. Data analysis was performed using the software environment R (R Core Team, 2018) and the software RStudio (RStudio Team, 2016). We first identified and excluded from further analyses those among the 26 participants exhibiting particularly low accuracy in their responses. The exclusion criterion was set so that the mean accuracy would not be smaller than the group average minus 2 standard deviations. 1 participant with only 39% correct answers was excluded. In our analyses, classical t-tests and anovas were used to compare average values. When focusing on invalid problems, we used the R package lme4 to model reasoning time values with a mixed-effect linear model, and accuracy with a mixed-effect logistic regression using binomial functions. Note that, below, we systematically report corrected sample standard deviation.

4.2 Results

We assessed the effect of counterexample density on invalid problems (valid problems do not present any counterexample). Overall, participants answered correctly to 66.0 $\pm$ 9.7 % of the invalid scanning problems, in 4.48 $\pm$ 0.43 secs. On average, for the problems presented with a diagram of low counterexample density, participants answered 64.5 $\pm$ 9.8 % correctly in 4.69 $\pm$ 0.42 secs. For the problems presented with a diagram of high counterexample density, participants answered 67.6 $\pm$ 9.6 % correctly in 4.30 $\pm$ 0.44 secs (Figure 8). While counterexample density did not significantly affect participants' accuracy (paired t-test: $t(24) = 1.5$, $p = 0.15$), we verified that it impacted their reasoning time (paired t-test: $t(24) = 2.47$ $p = 0.02$). To verify that this effect was not due to learning effects throughout the experiment, we then evaluated a mixed-effects linear model of reasoning time with the diagram condition (low/high counterexample density) and the order of presentation (to account for any learning effect throughout the experiment) as fixed effects. Note that we did not include the problems in this model after verifying that this factor did not have any significant effect on reasoning time (ANOVA: $F(1,11) = 1.23$, $p = 0.11$). Participants were modeled as a random effect. This model

revealed significant main effects of both the diagram condition ($F(1,21) = 7.03$, $p = 0.015$) and the order of presentation ($F(1,21) = 16.2$, $p < 0.001$), but no interaction between these two factors ($p = 0.88$). Thus, participants responded faster for the problems presented with a diagram of high counterexample density.

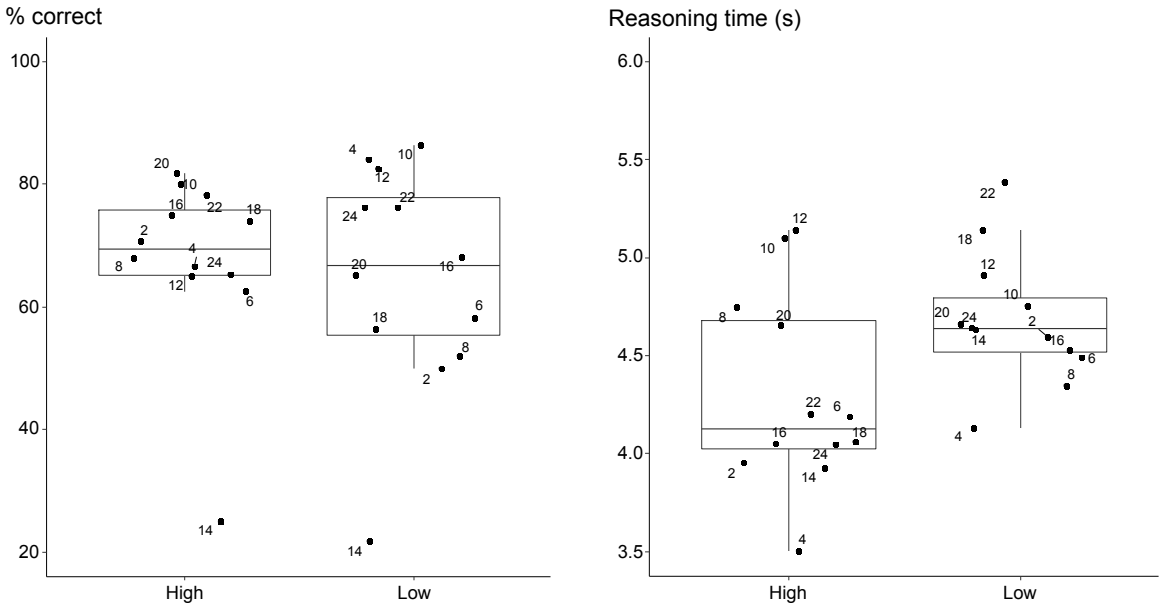


Figure 8. Percentages of correct responses and reasoning times (s) for the 12 invalid scanning problems of Experiments 2 in the two diagram conditions of "high" and "low" counterexample density.

4.3 Discussion

This experiment indicated that increasing the counterexample density of the diagram improved the reasoning performance on invalid problems, which was to be expected if a search for counterexamples through scanning operations underlies the reasoning. It also suggested that scanning operations are sensitive to metric aspects of the diagram, since the two diagrams of low and high counterexample density associated to each invalid problem were always topologically equivalent but metrically different.

5 Experiment 3

Another possible way to manipulate the difficulty of finding counterexamples in the diagram is in terms of *counterexample distance*. We define counterexample distance as the distance between the diagram and the "closest" counterexample (see Figure 9). Distance is measured here in terms of the traditional geometric transformations of translation, rotation, and uniform scaling as applied to the object of the considered scanning problem (e.g., line L in Figure 9). In this third experiment, we thus evaluated the effect of counterexample distance on adults' performance in scanning problems. In the cases of small counterexample distance, we expect participants to perform better.

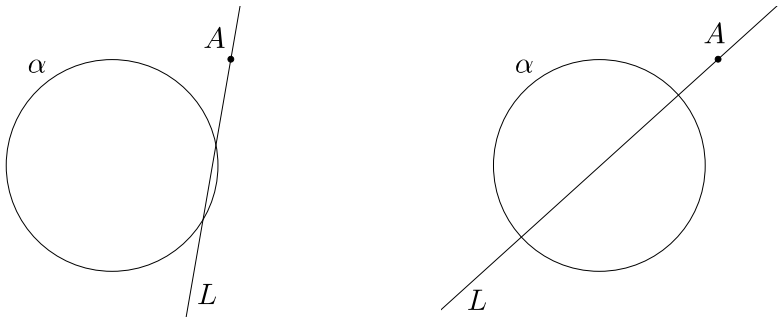


Figure 9. Two possible diagrams in the case of inference I_2. When considering a scanning problem with line L, then (a) is a diagram where L stands *close* to the closest counterexample and (b) is a diagram where L stands *far* from the closest counterexample.

5.1 Method

5.1.1 Design, Materials, Procedure, and Statistical Analysis. In this experiment, the participants were tested on the same 24 scanning problems as in our second experiment (see Table S5 and S6, sup. mat.). The procedure and statistical analysis were also the same as in the previous experiment, with the only exception that the images for the invalid problems were here manipulated in terms of counterexample distance. In other words, the 12 invalid problems were seen once under a configuration close to the closest counterexample, and once under a configuration far from the closest counterexample (see Table S6, sup. mat.).

5.1.2 Participants. We recruited a new group of 33 French speaking adults (15 females, age range: 19-54, mean: 31.4) on the Amazon Mechanical Turk platform with the same worker qualifications as before. We made sure that these participants did not already take part in the previous experiment. If they carried out the test in its entirety, they received $2 for their participation. According to the Declaration of Helsinki (2013), all participants provided informed consent prior to the experiment.

5.2 Results

To assess the effect of counterexample distance, we again looked at invalid problems since valid problems do not present any counterexample. Overall, the participants answered correctly to 68.9 $\pm$ 8.2 % of the invalid problems, in 4.46 $\pm$ 0.39 secs. On average, when the distance to the closest counterexample(s) was small, participants answered 73.5 $\pm$ 7.8 % correctly in 4.48 $\pm$ 0.39 secs. When the distance to the closest counterexample(s) was large, participants answered 64.4 $\pm$ 8.5 % correctly in 4.44 $\pm$ 0.40 secs (see Figure 10). This time, no significant effect of the counterexample distance was found on reasoning times (paired t-test: $t(32) = 0.41$, $p = 0.69$). However, a difference on accuracy was observed between the two diagram conditions (paired t-test: $t(32) = 2.80$, $p = 0.009$). To verify that this effect was not due to learning effects throughout the task, we performed a logistic regression with the diagram condition (close/far) and the order of presentation (first/second) as fixed effects and participants as random effect. This model revealed a significant effect of the diagram condition ($z = 2.15$, $p = 0.03$), and no effect of the order of presentation ($z = 0.57$, $p = 0.57$). These results thus revealed that participants responded significantly more accurately under the condition “close”.

5.3 Discussion

This experiment suggested that decreasing the distance between the object to which the reasoning question applied and the closest counterexample in the diagram improved the performance on the invalid problems. Similarly to Experiment 2, this was to be expected if a search for counterexamples through scanning operations underlies the reasoning. As in Experiment 2, the manipulation in terms of counterexample distance was a metric one. Experiment 3 thus provides further evidence that scanning operations are sensitive to metric aspects of the dia-

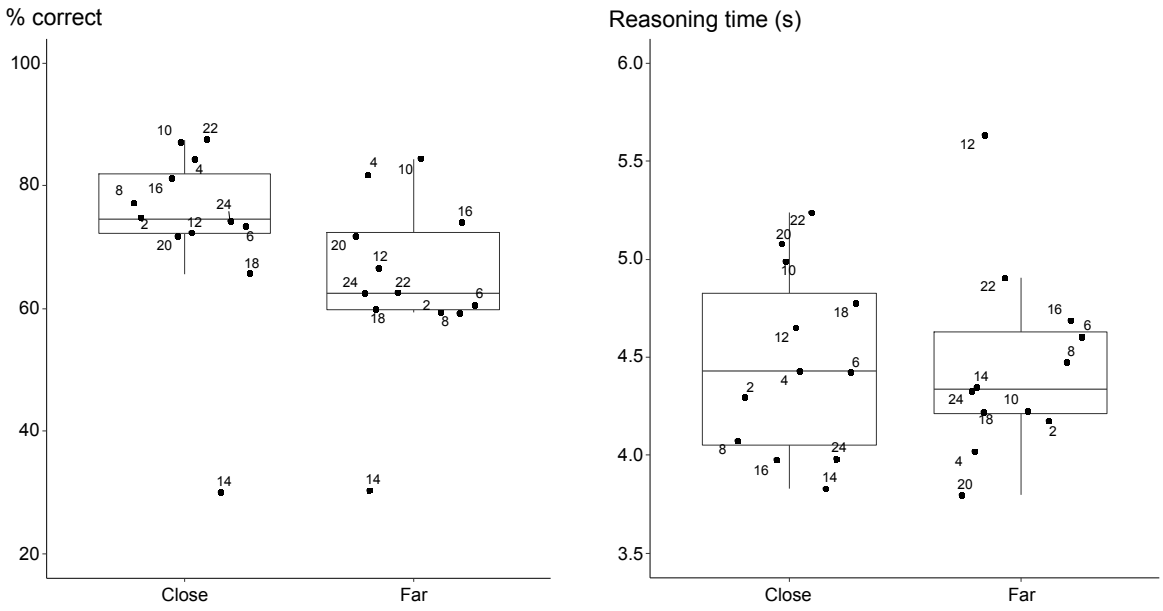


Figure 10. Percentages of correct responses and reasoning times (s) for the 12 invalid scanning problems of Experiments 3 in the two diagram conditions "close" and "far".

gram. We also observed here again that people can reach a high level of accuracy on scanning operations under a relatively short time constraint.

Finally, we note that manipulating counterexample distance affected the accuracy while manipulating counterexample density in Experiment 2 affected the reasoning time. This is compatible with the further hypothesis that scanning operations are performed locally and in parallel. Indeed, if scanning operations proceed locally, in a limited range around the initial position of the object, a counterexample is more likely to be hit if the object stands closer to the closest counterexample, thus resulting in a better accuracy. In addition, if the results of multiple scanning operations are evaluated in parallel, then between two configurations with the same relative distance to the closest counterexample, a counterexample is more likely to be hit faster in the case of high counterexample density, thus resulting in a shorter reasoning time.

6 General Discussion

In this study, we exposed adult participants to various diagram-based geometric inferences. In Experiment 1, we verified that they were able to carry out such inferences independently of any math training, and we revealed counterexample search as a potential mechanism of infer-

ential validity evaluation. On the basis of the diagrams that subjects produced, we hypothesized that the search for counterexamples proceeds through scanning operations that consist in varying mentally certain objects in the diagram, under the constraints expressed by the premisses, and with the objective of falsifying the conclusion. The results of Experiments 2 and 3 then indicated that scanning operations were sensitive to two kinds of metric manipulations of the diagram, as increasing the counterexample density or decreasing the counterexample distance in the diagram significantly improved subjects' reasoning performance.

First, we note that our participants almost always produced a diagram that correctly depicted the premisses. This suggests that, in our experiment, even naive reasoners were able to translate the information contained in sentences into a drawing. Noticeably, we observed some topological preferences among participants in diagram production (Table S2, supplementary materials). This result is in line with a phenomenon already described in spatial reasoning research (Ragni & Knauff, 2013). However, contrary to what has been previously observed in the production of quadrilaterals (Koedinger, 1998), the diagrams produced were not overly specific metrically.

Moreover, subjects who were able to produce a correct representation of the premisses and correctly judged an inference as invalid were almost always able to produce a counterexample to it. However, this differs from the results of most studies on reasoning and counterexample search in mathematics. For example, Koedinger (1998) has investigated the capacity of high-school students following a geometry course to reason about kites—quadrilateral figures *ABCD* where *AB* and *AD* are congruent to *CB* and *CD*, respectively. Students were asked to produce diagrams and formulate potential conjectures (e.g., about the diagonals), and then to find a proof or a counterexample to the formulated conjectures. Koedinger observed that students had trouble understanding the proper dialectic between deductive proofs, examples, and counterexamples. Similarly, when mathematics educators investigated how students use and understand counterexamples in various mathematical areas such as geometry and algebra (Buchbinder & Zaslavsky, 2019; Zaslavsky & Ron, 1998), number theory (Alcock & Inglis, 2008; Zazkis & Chernoff, 2008), and analysis (Ko & Knuth, 2009; Weber, 2009), they showed that one of students' main difficulties was to understand the proper role of examples and counterexamples in

proving and refuting mathematical claims. In our first experiment, participants from various backgrounds exhibited a remarkably high performance in our task, thus suggesting (1) a certain mastery of the notions of examples and counterexamples in the context of diagram-based geometric reasoning, and (2) that producing counterexamples is not reserved to math experts. The difference with the above mentioned studies may simply be due to the fact that we never explicitly asked for examples or counterexamples, but simply asked for illustrations.

Another well known context in which people have difficulties to find counterexamples is the Wason 's (1968) selection task. In this case, reasoners' capacity to identify potential counterexamples can increase drastically when the reasoning problem is formulated with concrete content (see, e.g., Griggs & Cox, 1982; Johnson-Laird, Legrenzi, & Legrenzi, 1972). Although the reasoning problems considered here concern abstract geometric objects and relations, the presence of diagrams supporting participants' reasoning could have helped them representing the abstract content of the problems in a more concrete way, thus making it easier to identify counterexamples. Alternatively, some researchers (see, e.g., Cheng & Holyoak, 1985; Stenning & van Lambalgen, 2008) have also argued that the difficulty of the Wason's selection task depends on whether the rule to be evaluated is interpreted in a descriptive or a deontic way. A descriptive interpretation means that the rule is interpreted as a material implication—e.g., "if a card has a vowel on one side, then it has an even number on the other"—while a deontic interpretation means that the rule is interpreted as involving a modal component saying what one can, should, or must do—e.g., "if one is to drink alcohol, then one must be over eighteen". It is possible that our participants have interpreted our geometric reasoning tasks as asking what can and cannot be drawn in a given geometric situation, thus introducing such a modal component. For instance, one may interpret inference I_1 as asking whether one can draw a line that goes through a point inside a circle without intersecting the circle. What facilitates counterexample identification in the Wason's selection task may then explain the rather good performance in counterexample production observed in Experiment 1. Further work would be necessary to determine whether performance on geometric reasoning is affected by varying the content of geometric problems along an abstract-concrete scale and/or by manipulating possible descriptive vs deontic interpretation of the task.

In the context of logical reasoning, two studies in the mental model tradition (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird & Hasson, 2003) have shown that people can construct counterexamples to refute inferences, but only to a certain extent. In a task where participants were asked to evaluate syllogistic inferences and to help themselves by constructing pictorial representations of the premisses using cut-out shapes, Bucciarelli and Johnson-Laird (1999) reported that participants engaged in the construction of alternative models for invalid inferences in only 48% of the cases, but noted that participants may have in some trials constructed alternative models mentally without externalizing them. In another task where participants were asked to evaluate sentential inferences and to write down a justification for their answers, Johnson-Laird and Hasson (2003) reported that participants used counterexamples as a justification for judging an inference as invalid in only 51% of the cases. Here again, those percentages are significantly lower than the one we observed in our first experiment. We can imagine that diagram-based geometric reasoning offers a less diverse variety of strategies to refute inferences in comparison to syllogistic reasoning (Bucciarelli & Johnson-Laird, 1999) and sentential reasoning (Johnson-Laird & Hasson, 2003). Our results suggest that the diagram supports a strategy that consists in decomposing the complex task of evaluating the validity of a geometric inference into scanning operations that proceed by varying mentally certain objects in the diagram while keeping the other objects fixed. We saw in Experiment 1 that subjects were more likely to carry out scanning operations on the objects present in the conclusion as compared to those only present in the premisses. This would suggest that the search procedure for counterexamples involves strategic decisions regarding the scanning operation(s) to be carried out. Characterizing these strategic decisions requires further investigation, for instance by recording how they use paper and pencil and/or by asking them to externalize their reasoning process through talk aloud protocols. This could also provide information on the order in which scanning operations are carried out, which may reveal preference patterns in the ordering of operations as previously observed in geometric analogy tasks (Novick & Tversky, 1987).

As mentioned earlier, the scanning operations we described in this study are different from the image scanning studied by Kosslyn et al. (1978) in that they do not concern static visual images but rather the possibilities to place one or more geometric objects, under certain con-

straints, in a geometric diagram. They are also different from the transformations on mental models investigated and theorized in the mental model theory approach to spatial relational reasoning (Goodwin & Johnson-Laird, 2005; Knauff, 2013; Ragni & Knauff, 2013) in that they are sensitive to metric information while mental models and the transformations thereof are only sensitive to the abstract or qualitative information provided in the premisses. The precise cognitive nature of the scanning operations though remains unknown. Our results are compatible with the hypothesis that scanning operations rely on object-based mental transformations (Zacks & Michelon, 2005) such as mental rotation (Shepard & Cooper, 1982; Shepard & Metzler, 1971), mental translation (Larsen & Bundesen, 1998), and mental uniform scaling (Besner & Coltheart, 1976; Bundesen & Larsen, 1975) which are all sensitive to metric information. Other hypotheses concern a potential role for visual routines (Ullman, 1984) and/or mental simulation of a physical kind (Hart et al., 2018). Further investigations will be necessary to distinguish between these hypotheses.

Our study focused on individual diagram-based geometric inferences, but many questions remain about the integration of such inferences in more complex geometric proofs of the kind found in Euclid's *Elements*. These proofs proceed through a combination of text and diagram, where the text mostly supports inferences about metric information, while the diagram supports inferences about topological information (Manders, 2008; Mumma, 2006). Thus one question is to evaluate the impact of the metric versus topological content on the reasoning process. The comparison between the high performance that our participants reached on topological inferences and the difficulties exhibited by students on metric inferences about kites in Koedinger's (1998) study may suggest that this distinction matters for counterexample search. Moreover, when Koedinger and Anderson (1990) investigated and modeled geometry proof search, they noticed that different types of inferences were treated differently in the abstract planning process—e.g., experts were more likely to skip algebraic inferences. Consequently, building on the work by Koedinger and Anderson (1990), and in particular examining whether text-based and diagram-based inferences are processed differently, could eventually shed light on the cognitive processes underlying geometric proof search where inferences concern both metric and topological information as in Euclid's proofs.

Finally, geometry is an abstract theory of space, but points, lines, circles and their relations are commonly used in concrete contexts when people reason with different kinds of visuospatial displays (diagrams, graphs, maps) or communicate and think in navigational contexts (Hegarty & Stull, 2012; Newcombe, 2018; Tversky, 2005). For instance, deductive geometric reasoning is essential to infer information that goes beyond what is depicted in a map. This is often necessary in route planning when one is wondering, say, whether a location can be reached without crossing a river or a mountain, or without having to enter into certain areas. As another example, let assume that you are at a position (or point) inside the Paris ring (i.e., a circle). Can you walk along a straight path (or line) indefinitely without crossing the ring? No. Now what if you are outside the Paris ring? Yes, and it is enough to think about a road between Bordeaux and Lyon to invalidate the above conclusion. The present study and its focus on counterexample search may thus provide a rather new approach to the question of deductive reasoning with topological relations between spatial entities such as positions, paths, roads, borders, regions, etc. More generally, the two notions of counterexample density and counterexample distance that we introduced in this study, and that affected reasoning performance, could ultimately prove useful to study other forms of reasoning. Indeed, our capacity to recognize any form of deduction as invalid could plausibly depend on the ratio of counterexamples in our semantic representation of the domain of reasoning or on the distance between the closest counterexample and the situation initially considered. Thus, studying the psychology of geometric reasoning is not only necessary to better understand how we think and reason about the geometric properties of our spatial environment and our spatial artefacts, it also provides a privileged context to investigate counterexample search in deductive reasoning.

Acknowledgements

We are grateful to Valeria Giardino, Véronique Izard, Jean Paul Van Bendegem, and David Waszek for reading through a pre-final version of this manuscript. We are also thankful to two anonymous reviewers of *Cognitive Science* for their comments and suggestions.

Funding

YH carried out the present work while holding a postdoctoral fellowship from the Research Foundation - Flanders (FWO). This research has also been supported by a postdoctoral fellowship attributed by the Fyssen Foundation to MA.

Conflict of interest

None.

7 References

Alcock, L., & Inglis, M. (2008). Doctoral students' use of examples in evaluating and proving conjectures. *Educational Studies in Mathematics*, *69*(2), 111–129.

Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., & Dehaene, S. (2017). The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLoS Computational Biology*, *13*(1), e1005273.

Avigad, J., Dean, E., & Mumma, J. (2009). A formal system for Euclid's *Elements*. *The Review of Symbolic Logic*, *2*(4), 700–768.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*(6), 372–378.

Besner, D., & Coltheart, M. (1976). Mental size scaling examined. *Memory & Cognition*, *4*(5), 525–531.

Booth, J. L., & Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. *British Journal of Educational Psychology*, *82*(3), 492–511.

Braine, M. D., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*(3), 247–303.

Buchbinder, O., & Zaslavsky, O. (2019). Strengths and inconsistencies in students' understanding of the roles of examples in proving. *The Journal of Mathematical Behavior*, *53*, 129–147.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 214–220.

Burgess, N. (2008). Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, *1124*(1), 77–97.

Byrne, R., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, *28*(5), 564–575.

Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression

of inferences. *Journal of Memory and Language*, *40*(3), 347–373.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an amazonian indigene group. *Science*, *311*(5759), 381–384.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

De Soto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, *2*(4), 513–521.

Dillon, M. R., Huang, Y., & Spelke, E. S. (2013). Core foundations of abstract geometry. *Proceedings of the National Academy of Sciences*, *110*(35), 14191–14195.

Euclid. (1959). *Elements* (D. Densmore, Ed.). New York: Dover Books. (published as *Euclid's Elements: all Thirteen Books Complete in One Volume*, and translated by T. L. Heath)

Evans, J. S. B., Newstead, S. E., & Byrne, R. M. (1993). *Human reasoning: The psychology of deduction*. Hove: Lawrence Erlbaum.

Giaquinto, M. (2011). Crossing curves: A limit to the use of diagrams in proofs. *Philosophia Mathematica*, *19*(3), 281–307.

Goodwin, G., & Johnson-Laird, P. (2005). Reasoning about relations. *Psychological Review*, *112*(2), 468–493.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, *73*(3), 407–420.

Hart, Y., Dillon, M. R., Marantan, A., Cardenas, A. L., Spelke, E., & Mahadevan, L. (2018). The statistical shape of geometric reasoning. *Scientific Reports*, *8*(1), 12906.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1084–1102.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Hegarty, M., & Stull, A. T. (2012). Visuospatial thinking. In K. J. Holyoak & J. Morrison

Robert (Eds.), *The oxford handbook of thinking and reasoning* (pp. 606–630). Oxford: Oxford University Press.

Heiser, J., & Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. *Cognitive Science*, *30*(3), 581–592.

Izard, V., O'Donnell, E., & Spelke, E. S. (2014). Reading angles in maps. *Child Development*, *85*(1), 237–249.

Izard, V., Pica, P., Spelke, E., & Dehaene, S. (2011). Flexible intuitions of Euclidean geometry in an Amazonian indigene group. *Proceedings of the National Academy of Sciences*, *108*(24), 9782–9787.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford: Oxford University Press.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250.

Johnson-Laird, P. N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition*, *31*(7), 1105–1113.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, *63*(3), 395–400.

Kao, Y., Douglass, S., Fincham, J., & Anderson, J. (2008). Traveling the second bridge: Using fMRI to assess an ACT-R model of geometry proof. Research report, Department of Psychology, Carnegie Mellon University.

Kline, M. (1972). *Mathematical thought from ancient to modern times*. New York: Oxford University Press.

Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation*, *1*(3), 261–290.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.

Knauff, M., Strube, G., Jola, C., Rauh, R., & Schlieder, C. (2004). The psychological validity of qualitative spatial reasoning in one dimension. *Spatial Cognition and Computation*, *4*(2), 167–188.

Ko, Y.-Y., & Knuth, E. (2009). Undergraduate mathematics majors' writing performance pro-

ducing proofs and counterexamples about continuous functions. *The Journal of Mathematical Behavior*, *28*(1), 68–77.

Koedinger, K. R. (1991). *Tutoring concepts, percepts, and rules in geometry problem solving* (Unpublished doctoral dissertation). Carnegie Mellon University.

Koedinger, K. R. (1998). Conjecturing and argumentation in high school-geometry students. In R. Lehrer & D. Chazan (Eds.), *New directions in the teaching and learning of geometry* (pp. 319–347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Koedinger, K. R., & Anderson, J. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, *14*(4), 511–550.

Koedinger, K. R., & Anderson, J. (1993). Effective use of intelligent software in high school math classrooms. In *Artificial intelligence in education: Proceedings of the world conference on artificial intelligence in education* (pp. 241–248). Charlottesville, VA: AACE.

Koedinger, K. R., & Terao, A. (2002). A cognitive task analysis of using pictures to support pre-algebraic reasoning. In C. Schunn & W. Gray (Eds.), *Proceedings of the twenty-fourth annual conference of the cognitive science society* (Vol. 24, pp. 542–547). Mawah, NJ: Lawrence Erlbaum Associates, Inc.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(1), 47–60.

Landy, D., & Goldstone, R. L. (2007a). Formal notations are diagrams: Evidence from a production task. *Memory & Cognition*, *35*(8), 2033–2040.

Landy, D., & Goldstone, R. L. (2007b). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 720–733.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10*(6), e0130834.

Larsen, A., & Bundesen, C. (1998). Effects of spatial separation in visual pattern matching: Evidence on the role of mental translation. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 719–731.

Macbeth, D. (2010). Diagrammatic reasoning in Euclid's *Elements*. In B. Van Kerkhove, J. De Vuyst, & J. P. Van Bendegem (Eds.), *Philosophical perspectives on mathematical practice.* London: College Publications.

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108–114.

Manders, K. (2008). The Euclidean diagram. In P. Mancosu (Ed.), *Philosophy of mathematical practice* (pp. 80–133). Oxford: Oxford University Press.

Matsuda, N., & VanLehn, K. (2004). GRAMY: A geometry theorem prover capable of construction. *Journal of Automated Reasoning*, *32*(1), 3–33.

Miller, N. (2007). *Euclid and his twentieth century rivals: Diagrams in the logic of euclidean geometry*. Stanford: CSLI Publications.

Mueller, I. (1981). *Philosophy of mathematics and deductive structure in euclid's elements*. Cambridge, MA: The MIT Press.

Mumma, J. (2006). *Intuition formalized: Ancient and modern methods of proof in elementary geometry* (Unpublished doctoral dissertation). Carnegie Mellon University.

Mumma, J. (2012). Constructive geometrical reasoning and diagrams. *Synthese*, *186*(1), 103–119.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Netz, R. (1999). *The shaping of deduction in greek mathematics: A study in cognitive history*. Cambridge: Cambridge University Press.

Newcombe, N. (2018). Three kinds of spatial cognition. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience (fourth edition)* (pp. 1–31). Hoboken, NJ: John Wiley & Sons, Inc.

Novick, L. R., & Tversky, B. (1987). Cognitive constraints on ordering operations: The case of geometric analogies. *Journal of Experimental Psychology: General*, *116*(1), 50–67.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human*

reasoning. Oxford: Oxford University Press.

Panza, M. (2012). The twofold role of diagrams in Euclid's plane geometry. *Synthese*, *186*(1), 55–102.

Pedone, R., Hummel, J. E., & Holyoak, K. J. (2001). The use of diagrams in analogical problem solving. *Memory & Cognition*, *29*(2), 214–221.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*(3), 533–566.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

Rips, L. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: The MIT Press.

Ritter, S., Anderson, J., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255.

RStudio Team. (2016). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from `http://www.rstudio.com/`

Shah, P., & Miyake, A. (2005). *The cambridge handbook of visuospatial thinking*. Cambridge: Cambridge University Press.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, *19*(1), 97–140.

Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, *34*(2), 109–159.

Tversky, B. (2005). Visuospatial reasoning. In K. J. Holyoak & J. Morrison Robert (Eds.), *The cambridge handbook of thinking and reasoning* (pp. 209–240). Cambridge: Cambridge University Press.

Ullman, S. (1984). Visual routines. *Cognition*, *18*(1-3), 97–159.

Van der Henst, J.-B. (2002). Mental model theory versus the inference rule approach in relational reasoning. *Thinking & Reasoning*, *8*(3), 193–203.

Wang, R. F., & Spelke, E. S. (2002). Human spatial representation: Insights from animals. *Trends in Cognitive Sciences*, *6*(9), 376–382.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Weber, K. (2009). How syntactic reasoners can develop understanding, evaluate conjectures, and generate counterexamples in advanced mathematics. *The Journal of Mathematical Behavior*, *28*(2-3), 200–208.

Zacks, J. M., & Michelon, P. (2005). Transformations of visuospatial images. *Behavioral and Cognitive Neuroscience Reviews*, *4*(2), 96–118.

Zaslavsky, O., & Ron, G. (1998). Students' understandings of the role of counter-examples. In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22nd conference of the international group for the psychology of mathematics education* (Vol. 4, pp. 225–232). Stellenbosch: Program Committee of the 22nd PME Conference.

Zazkis, R., & Chernoff, E. J. (2008). What makes a counterexample exemplary? *Educational Studies in Mathematics*, *68*(3), 195–208.