Preprint

The Long-Term Effect of Relational Information in Classification Learning

Fabien Mathy
Université de Franche-Comté

This study examines the long-term effect of mutual information in the learning of Shepardian classifications. Mutual information is a measure of the complexity of the relationship between features because it quantifies how the features relate to each other. For instance, in various categorization models, Type VI concepts –originally studied by Shepard, Hovland, and Jenkins (1961)– are unanimously judged to be the most complex kind of 3D Boolean concepts. This has been largely confirmed by empirical data. Yet, it is apparently inconsistent with the fact that this concept entails the greatest amount of mutual information of all the 3D Boolean concepts. The present study was aimed at verifying whether individuals can use relational information, in the long run, to devise easier strategies for category learning. Subject performance was measured repeatedly for one hour on either successive Type VI concepts (using different features between problems) or successive Type IV concepts. The results showed that shortly after the second problem, Type VI concepts became easier to learn than Type IV ones. The gap between the mean per-problem error rates of the two concepts continued to increase as the number of problems increased. Two other experiments tended to confirm this trend. The discussion brings up the idea of combining different metrics in categorization models in order to include every possible way for subjects to simplify the categorization process.

The aim of this paper is to show that certain concepts (like the Type VI classification, described later) entail relational features that can facilitate the learning of these concepts in the long term. Mutual information is investigated here as a generic measure of relational information. The relatedness between variables is not supported by most categorization models which see Type VI concepts as the most difficult kind. However, the three experiments presented reveal that individuals do use relational information in the long run and eventually find Type VI concepts easier. I discuss the fact that classification models tend to rely on single metrics that do not take into account the multiple strategies that individuals might use simultaneously.

Mutual information is a generic measure of relational information or redundancy between variables which has been developed within the context of information theory (Shannon, 1948; Garner, 1962). In communicational systems, mutual information can be understood as the amount of transmitted information between input and output. If the input is correlated to the output, it means that all information has been transmitted without any loss of information. Information theory was extensively used in the 1950s and 1960s, specially for measuring the maximal amount of information that can be transmitted by subjects without error (the reader will find introductory presentations of this approach in Attneave, 1959; Coombs, Dawes, & Tversky, 1970). In an early work (Miller, 1956, one of the most cited of the *Psychological Review*, reprinted in Miller, 1994), Miller related several experiments on absolute judgments which tended to show that the channel capacity of subjects was limited to about 7 alternative stimuli (about $log_2(7) = 2.8$ bits), after which the information could not be transmitted without errors (the subjects could not perfectly match the set of responses to the set of stimuli). Simultaneously, Miller began contributing to the decline of information theory by making a distinction between bits of information and chunks of information. His argument was that the memory span is limited in the number of chunks (7), but not in the number of bits, thus showing that short-term memory does not fit a model of channel capacity [1]. Luce (2003) gives an interesting historical account of the shifting away from information theory in psychology after the 1960s. Despite this shortcoming, the channel capacity model is still used in absolute identification experiments (Houtsma, 1983; Mori, 1998) or in other domains such as intelligence measurement (Lehrl & Fischer, 1990). Mutual information on more than two variables is more rarely in-

The author would like to thank David Fass who kindly introduced him to the notion of mutual information. Special thanks go to David Fass and Jacob Feldman (Visual Cognition Lab, Center for Cognitive Science, Rutgers University, NJ) for their helpful comments. This research was supported in part by a postdoctoral research grant from the Fyssen Foundation in 2005. Correspondence concerning this article should be addressed to Fabien Mathy, Université de Franche-Comté - UFR SLHS, 30-32 rue Mégevand - 25030 Besançon Cedex, France. Email: fabien.mathy@univ-fcomte.fr

[1] Miller's argument was as follows: the processing of a numerical digit requires 3.3 bits because the best encoding scheme for the 10 digits needs a combination of 3.3 binary digits ($log2(10) = 3.3$). If the short-term memory capacity extends to 7 numerical digits, the sum of information that can be transmitted in short-term memory is $7 \times 3.3 = 23$. This limitation should then be applied to other items. For instance, because words require around 10 bits of coding, short-term memory should be limited to around 2.3 words. However, as reported by Miller, the capacity is rather limited to 6 or 7 words.

vestigated in psychology (cf. Fass, 2006, who gives a comprehensive review and proposes direct applications in human causal learning). The aim of this paper is to demonstrate that mutual information is a very useful tool indicating how sets of variables are related to one another. Some detailed explanations for computing mutual information are given below and in the appendix.

First, two examples of concepts that might be learned using relational information (including the Type VI concept) are given. This is followed by an overview of a subset of classification models which unanimously claim that Type VI concepts are the most complex kinds of three dimensional concepts, a claim that the mutual information metric refutes.

Simplicity of Relational Concepts: XOR and Type VI

A couple of strategies based on relational information might be used to learn the two concepts presented in Table 1. Concepts can be viewed as categorization situations restricted to two categories only (the category of positive examples versus the category of negative examples), no matter the number of input dimensions (some features). This article deals with artificial concepts created by arbitrarily assigning positive membership to a list of stimuli. Artificial concepts are specially adapted to testing cognitive models because the difficulty can be manipulated using different mappings between the critical features and the categories. The first concept presented in the table is the exclusive OR (generally called XOR in logic) and the second is called Type VI (originally studied and named by Shepard, Hovland, & Jenkins, 1961). All stimuli have been generated using combinations of black and white balls, but other features could have been used as well. Here, each group of horizontally aligned balls represents a single stimulus. Using N balls in a row and two colors, 2^N stimuli can be built. The categories are chosen using + (for positive examples) and - symbols (for negative examples). In the XOR, as in any other concept, the learner needs to differentiate the positive examples from the negative ones. If one feature was characteristic of the class (that would be the case if, for instance, all positive stimuli had a first ball colored in white), subjects could use a concise rule such as "if the first ball is white, then the stimulus is positive". Because no feature is characteristic of the class in the XOR, it first appears that the learner must memorize that "if the first ball is white and the second is black, or if the first ball is black and the second is white, then the stimulus is positive". In this case, the rule strictly corresponds to the list of the positive stimuli. However, by using relational information in the XOR, one can notice that the stimuli are simply positive if the balls have different colors, which might be more economical. Here, the features of one ball or the other is not diagnostic when one notices that the relation "different" between the two balls transcends the particular feature values. Similarly, the Type VI concept has apparently no critical feature, so the four positive examples seem to have to be learned by rote. However, using relational information one might notice that the second and third balls have different colors when the first is white, whereas they have identical colors when the first is black. Such higher order rules might facilitate the learning of the XOR and the Type VI classifications. Relational information is even more powerful in Type VI: what was true for the first ball is also true for the second and the third (for instance, if the third ball is white, the stimulus is positive if the other two balls have different colors, and so forth); also, relational information comes with a numerical facilitation as one might notice that the stimuli are positive only if the number of black balls is odd (i.e., the stimulus is positive if it contains one or three black balls). However, mutual information is not always that readily apparent. For instance, using more complex stimuli in which dimensions are incommensurate, mutual information needs to be used by subjects in a more complex manner instead of simply using same/different relationships. The aim of the paper is to show that mutual information can also allow strategies based on simple hierarchies of rules in some circumstances. These strategies can be used as a powerful abstraction process to reduce the apparent complexity in certain categorization tasks. The hypothesis here is that such higher order rules might be discovered as long as enough time is allocated to learning. We will briefly examine a subset of classification models which does not take such information into account.

Complexity of Type VI Concepts

In the categorization literature, the goal is usually to determine which model best predicts participant performance. To the best of my knowledge, none of the categorization models offer a metric of conceptual complexity which relies on relational information. The present study demonstrates that none of these models is able to predict the beneficial long-term effect of relational information on learning.

Figure 1 shows the six basic concept types with three Boolean dimensions and four positive examples (first studied by Shepard et al., 1961). Table 2 indicates how these six concepts are ranked, by five classical categorization models, according to complexity. The literature on categorization is mainly devoted to comparing exemplar models that use similarity as a metric in the psychological space (Estes, 1994; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994) and logical models that use compression to abstract the simplest categorization rules (Bourne, 1970; Bruner, Goodnow, & Austin, 1956; Bradmetz & Mathy, 2008; Feldman, 2000, 2006; Hovland, 1966; Lafond, Lacouture, & Mineau, 2007; Mathy & Bradmetz, 2004; Nosofsky, Palmeri, & McKinley, 1994; Vigo, 2006). Therefore, the opposition lies mainly between geometrical metrics and algorithmic metrics. Sim-

Table 1
The XOR and TYPE VI concepts

	Colors	$\mathcal{C}$	Stimuli
XOR	***w w***	-	○○
	w b	+	○●
	b w	+	●○
	b b	-	●●
TYPE VI	***w w w***	-	○○○
	w w b	+	○○●
	w b w	+	○●○
	w b b	-	○●●
	b w w	+	●○○
	b w b	-	●○●
	b b w	-	●●○
	b b b	+	●●●

Note. Column C indicates the category membership; *w* (white) and *b* (black) indicate the color of the balls used in the stimuli. Each group of horizontally aligned balls represents a single stimulus.

ply put, exemplar models predict that concept are easy when categories are homogeneous (the categories are homogeneous when the examples are very similar within categories), whereas logical models predict that concepts are easy when a simple rule can be used by subjects to categorize the examples. The five models in Table 2 are derived from these two classes. The first model, called the raw similarity model (RSM) computes the probability of categorizing each example, assuming that the classification of an example is determined by its similarity to the stored category exemplars (Medin & Schaffer, 1978; Nosofsky, 1984; some of the principles were developed by Luce, 1963 and Shepard, 1957). It is a raw similarity choice model, including no parameters. Exemplars form a context for computing similarities between a stimulus presented to the subject and each category exemplar, the exemplars being the psychological representatives of the corresponding concrete examples of a given concept. A simple distance function was used with a city-block metric (counting the number of different features between two stimuli), n the number of dimensions composing the stimuli (here, $n = 3$), and x_{ia} the value of stimulus i on dimension a.

$$d_{ij} = [\sum_{a=1}^{n} |x_{ia} - x_{ja}|] \tag{1}$$

The following exponential decay function was used to relate stimulus similarity to psychological distance (Nosofsky, 1986; Shepard, 1987):

$$\eta_{ij} = e^{-d_{ij}} \tag{2}$$

Given the total similarity of a stimulus i to all exemplars in categories X and Y, the probability of responding with category X was computed by Luce's choice rule:

$$P(X/i) = \frac{(\sum_{x \in X} \eta_{ix})}{(\sum_{x \in X} \eta_{ix}) + (\sum_{y \in Y} \eta_{iy})} \tag{3}$$

To obtain a measure of the complexity of the concepts, a single probability term was computed by taking the average of all $P(CorrectCategory/i)$. Nosofsky (1984) showed that such a model gives a wrong prediction of the complexity of Type II (ranked fifth instead of second in Table 2), meaning that Type II is more complex than expected. As demonstrated below, more accurate exemplar models need to allow for selective attention processes (Nosofsky, 1984; Kruschke, 1992).

The second model, called General Context Model (GCM), represents a more optimal context model because it implements several parameters, including (1) a sensitivity parameter interpreted as an overall parameter for scaling distances in the space, and (2) a parameter used for weighting the attention paid to each dimension (Nosofsky, 1984). The preceding distance function is augmented with the scale parameter c reflecting discriminability in the psychological space and $n = 3$ attention weight parameters (one per dimension) with $0 \leq w_a \leq 1$, and $\sum w_a = 1$ ($n - 1 = 3 - 1$ were free to vary).

$$d_{ij} = c[\sum_{a=1}^{n} w_a |x_{ia} - x_{ja}|] \tag{4}$$

This model is notably a better predictor of the simplicity of the Type II concept (Nosofsky, 1984). More powerful GCM is possible, but no extra parameters such as the bias or γ were necessary to obtain a correct ranking of the concept Types here.

Other models can be likened to rule-based models in which compression is used to reduce the total amount of information in a set of examples of a given category. In the third model, developed by Feldman (2000) (called DNF-F here, for Disjunctive Normal Form-Feldman), the subjective complexity of a set of positive examples is mapped to the length of the shortest logically equivalent propositional formula (called the minimal DNF). For instance, if three Boolean dimensions are distinguished as follows: shape (square, s, or not square, s'; the apostrophe denoting the negation of the feature s), size (large, l, or not large, l'), and color (blue, b, or not blue, b'), and a list of positive examples is

$$1 = sl'b',\ 2 = sl'b,$$

a propositional formula made of disjunctions of features (put in conjunctions) can be written

$$(s \wedge l' \wedge b') \vee (s \wedge l' \wedge b)$$

where conjunctions and disjunctions are respectively represented by the symbols $\wedge$ and $\vee$. The formula is more readable when written as follows: $sl'b' \vee sl'b$. This formula simply describes the list of positive examples in a Boolean form that can be used for testing whether an example is positive or not. This formula can however be rewritten more economically as sl' (since the color feature is not relevant), without losing any potential in testing examples. Because formulas cannot be easily reduced when categories are heterogeneous, the number of features in the shortest/compressed formula (i.e., the minimal DNF) can be seen as a complexity index. In contrast with exemplar models, in which categorization seems associative and automatic, rule-based models seem to operate on a more controlled process of abstraction of information, a process maybe in part supported by language [2]. The complexity index was used here to get the ranking of all Types.

The fourth model (called DT) is similar to the third, except that the reduction technique is based on static decision trees in which all decisions are made of identical orderings of the relevant dimension values (see Bradmetz & Mathy, 2008; Mathy & Bradmetz, 2004, or see Bryant, 1986 who gives an account of similar reduction techniques using ordered binary decision diagrams). The fifth model (called DNF-KV) roughly corresponds to Felman's, except that the reduction technique makes use of Karnaugh-Veitch diagrams (or equivalent computerized techniques) to increase the compressibility of some concepts (Bradmetz & Mathy, 2008; Lafond et al., 2007; Vigo, 2006). The ranking of all Types by all models is given in Table 2.

Note that in Table 2, Type VI is ranked by all models as the most complex concept, although some differences between other concepts are noticeable between models (the second part of Table 2 shows the correlations between the rankings). Type VI concepts are unanimously assessed as the most difficult ones, either because of the heterogeneity of categories resulting from the pattern of dissimilarities between examples (for exemplar models) or because of the incompressibility of the set of positive examples (in models assessing minimal DNFs). GCM give different rankings for Types III, IV and V, but these three Types are, however, generally considered of equal difficulty, as the patterns of probabilities are quite close for these concepts. GCM consequently give a ranking (1,2,4,4,4,6) similar to the two first rule-based models. Since the work of Shepard el al., the ranking often summarized as $I < II < (III, IV, V) < VI$ has unanimously been replicated in categorization tasks in which subjects made more errors when the Type number was higher (the most famous replications are those by Feldman, 2000, and Nosofsky, Gluck, et al., 1994).

However, the argument stressed in this paper is that learning strategies in classification tasks are not limited to computing similarities or formulating rules on the basis of some relevant features. Individuals might use other means of reducing the complexity of these tasks, one of which is relational complexity developed below. The experiments which follow aim at showing that some strategies based on relational complexity can be found in subjects when these tasks are given several times in a row. We predict that the use of relational complexity will distort the traditional pattern of complexity of the Shepardian Types predicted and observed in most studies. The goal here is specially to indicate the considerable conceptual power of relational complexity by reversing the classical prediction that Type VI is the most difficult $3D$ Boolean concept.

Relational Complexity and Mutual Information

None of the categorization models presented above are able to measure the relational information that might exist between variables. Reference must be made here to the notion of mutual information, which measures how features relate to each other. I will begin with an intuitive approach, but a more complex and general way of measuring the relatedness between variables using mutual information is given in the appendix. For the sake of simplicity, the notions of relational complexity and mutual information are developed using the exclusive disjunction (XOR) described in Figure 2, because the Type VI concept is merely an extension of the XOR function and its properties.

When examining the truth table of the XOR (Fig. 2), one can notice that the output (i.e., the class) is positive (that is, equal to one) whenever the input values are different (10 or 01). In this case, the relation ("different") between the values is more informative than the values themselves. This relational complexity is easily shown in a Bayesian network (Fig. 2), which indicates that variables 1 and 2 are two immediate causes of the variable "Class"[3] (cf. Glymour, 2001; Pearl, 2000). When more complex dimensions are used (circle vs square, and white vs black, instead of simple 0/1 codes), it is not possible to use differences between the values of different dimensions (e.g., a circle value cannot be judged differently from a white value) . However, there is a

[2] Exemplar models and rule-based models seem apparently opposed here, but they can be seen as complementary (Sloman, 1996) and can be implemented in hybrid models (Anderson & Betz, 2001), the hybrid position being supported by neurophysiological evidence (Smith & Grossman, 2008).

[3] A peculiar property of the XOR is that relational complexity is maximal, which means that the three variables can be permuted and any of them can act as the class, because as long as the other two are different, the third is equal to one. Therefore, three Bayes nets could be drawn from the truth table depicted in Fig. 2.

Table 2
Ranking of the difficulty of the six classification types investigated by Shepard, Hovland, & Jenkins (1961) in different categorization models, measures of agreement between rankings, and mutual information within each type

	TYPE						Correlations			
	1	2	3	4	5	6	*b*	*c*	*d*	*e*
a- RSM	1	5	3	2	4	6	.600	.447	.447	.552
b- GCM	1	2	4	3	5	6		.894*	.894*	.828*
c- DNF-F	1	2	4	4	4	6			1**	.849*
d- DT	1	2	4	4	4	6				.849*
e- DNF-KV	1	2.5	2.5	4	5	6				
Mut. Info.	0	0	-.12	.06	.18	1				

Note. RSM, raw similarity model, using city-block distance (Nosofsky, 1984); GCM, General Context Model using city-block distance, attention weight parameter and sensitivity parameter (Nosofsky, 1984); DNF-F, with minimal disjunctive normal forms generated using Feldman's heuristics (cf., Feldman, 2000); DT, basic sequential and static Decision Tree model (cf., Bradmetz & Mathy, 2008; Mathy & Bradmetz, 2004); DNF-KV, with disjunctive normal forms computed using the Karnaugh-Veitch diagrams or equivalent techniques (cf., Bradmetz & Mathy, 2008; Lafond et al., 2007; Vigo, 2006); Kendall's τ correlations, computed on ranks; ** significant at the 0.01 level; * significant at the 0.05 level; Mut. Info., Mutual information.

simple connection of relational information with rule-based models. For instance, the learner can notice that if the first variable is equal to 0, then the values of the category variable and the second variable are correlated, whereas this relation is reversed if the first variable is equal to 1. Concretely, if examples are black, circles are positive and squares are negative, whereas if examples are white, a reverse pattern occurs (i.e., circles are negative and squares are positive). The learner might find the relational structure more obvious by taking one stimulus as a referent object. Knowing that the black circle is positive, the learner might notice that the object which is neither black nor circle is also positive. Then, the learner might quickly understand that all other objects are negative. The rule would state something like "IF [black and circle] OR [NOT black and NOT circle], THEN +, else -), letting the learner memorize only two features. Again, the values are not important anymore, except that at least one stimulus must be taken as a referent object when dimensions have values other than 0s or 1s.

Figure 1 shows a decision tree for categorizing the examples of a Type VI concept. Looking carefully at this tree, one can see a series of oppositions between decisions, which results from the presence of relational information: the left and right subtrees (from the second level to the leaves) are exactly the same except that the categories are reversed (details are given in the figure notes). Therefore, the preceding rule ("IF (black and circle) OR (NOT black and NOT circle), THEN +, else -) which corretly applies to the big objects might be reversed for the small objects ("IF ... THEN -, else +). Type VI is therefore made of two reversed XOR structures.

It is hypothesized here that individuals can use this relational information to devise easier strategies (i.e., using simple reversed decisions) to learn categories.

Mutual information simply quantifies the relatedness between two or more variables. In two dimensions, for example, mutual information corresponds to the reduction in the uncertainty about one variable due to the knowledge of another variable (see Duda, Hart, & Stork, 2001, pp. 630-632). In this case, mutual information would be maximal (i.e., equal to 1) in the case of a perfect positive or negative correlation between the two variables. In three dimensions, mutual information is maximal (i.e., equal to 1) in the case of a perfect positive or negative correlation between two variables holding the 3rd variable constant, which is the case for the XOR structure (as demonstrated above). In four dimensions, mutual information is maximal (i.e., equal to 1) in the case there is an XOR structure between three variables when the fourth variable is fixed, which is the case for the Type VI structure (as demonstrated above). Therefore, Type VI also entails mutual information equal to 1. The appendix provides some additional necessary details to compute mutual information. When learning the Type VI concept, the learner might notice that if the first input value is equal to one, the category variable is equal to one if the other two input variables are correlated, whereas if the first input value is equal to zero, the category variable is equal to one if the other two input variables are inversely correlated. Again, this translates in reversing decisions from one face of the cube to the other.

Table 2 gives the mutual information for each of the six Shepardian classification types [4]. I intend here to give a simple and clear picture of the effect of mutual information on

[4] Note that mutual information can be negative. This happens

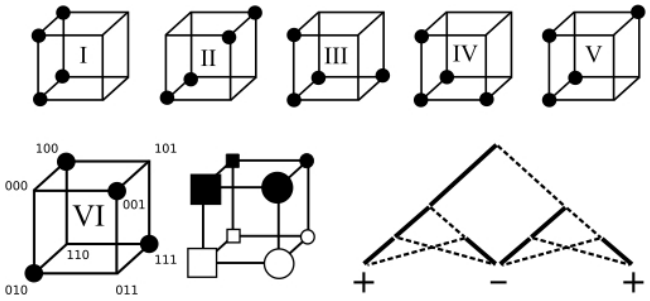
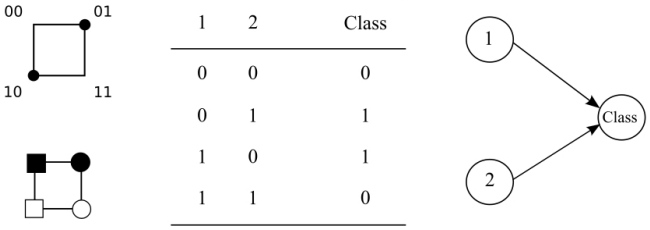


Figure 1. The six classification types studied by Shepard, Hovland, & Jenkins, 1961. *Note.* Positive examples are shown as black circles; negative examples are represented by empty vertices. There are only six possible conceptual structures in three Boolean dimensions with four positive examples. The other concepts are equivalent by rotation or mirror reflection. The seventh cube shows a set of possible stimuli using size, shape, and color as dimensions. This is followed by an appropriate decision tree for categorizing the examples in the Type VI concept. In the decision tree, the solid lines represent (from top to bottom) the values Big, Square, and White, respectively, whereas the dashed lines represent the values Small, Circle, and Black. For instance, taking three successive left branches (meaning: IF "Big, Square, and White") leads to a leaf representing a positive example (meaning: THEN "Category is +"); taking a right branch then two left branches (meaning: IF "Small, Square, and White") leads to a leaf representing a negative example (meaning: THEN "Category is -"). Note that the two subtrees (i.e., the two small decision trees following the first level) are equivalent except for their leaves which are reversed (the positive leaves are on the left for the first subtree whereas the positive leaves are on the right for the second one). This means that once the classification for the big examples is learned, subjects can apply a reversed classification for the small ones because of the presence of relational information.



1	2	Class
0	0	0
0	1	1
1	0	1
1	1	0

Figure 2. The exclusive disjunction: truth table and corresponding Bayesian network. *Note.* The truth table includes two input variables and one output variable. The output variable is called the class or the category in the categorization literature. The Bayesian network shows that input 1 and input 2 are independent. However, in this network, input 1 is NOT conditionally independent of input 2, GIVEN the class. Effectively, the value of input 2 is equal to the value of input 1 if the class is 0, whereas input 1 and input 2 are inversely correlated if the class is 1. The same properties would follow if we permuted the variables in the network. In other words, if any two variables are the same, the third is equal to zero, whereas if any two variables are different, the third is equal to one. Therefore, one can correctly classify the examples of an XOR by considering that the category is negative if the two input values are the same, and positive if not.

classification learning. To begin with a simple experiment, let us oppose Type VI with another 3D concept which has the smallest amount of positive mutual information, namely, a Type IV concept. Type II concepts have no mutual information, which would have made them suitable for a comparison, but because only two dimensions are relevant in this type of concept, it cannot simply be compared to Type VI (idem for Type I which has only one relevant dimension). Type V is also suitable for comparison, but it might have too much mutual information to be easily distinguished from Type VI. The first two experiments therefore aim at showing a clear difference between learning Type IV and Type VI in the long run, whereas all Types will be compared in Experiment 3.

Shepard et al., 1961 showed that when subjects are given successive problems of the same Type, more within-type transfer can be attributed to Type VI than to Types III, IV, and V (cf. Figure 6, p. 8). The problem is that 1) the individual curves for Types III, IV and V were not presented separately (Type IV was the easiest of Types III, IV and V with respect to the mean number of errors across the whole experiment, so Type IV might have a within-type transfer comparable to Type VI) and 2) the measures on Type IV were based on only two subjects (because each of the six subjects were assigned one of the three III, IV or V Types), so a complete counterbalancing was not achieved (Type IV was only learned first by one subject and last by the second, whereas Type VI was learned by four subjects at four different positions).

Therefore, 1) Types III, VI and V need to be studied separately in order to better equilibrate the comparison with Type VI. For that matter, in Experiments 1 and 2, Type IV will be individually compared to Type VI for some reasons developed below. In Experiment 3, the six Types will be studied individually but aggregated differently than in Shepard et al.'s analysis. 2) Reliability and generalizability of the results need to be tested with larger samples of subjects. As reported by Shepard el al. (p. 9), the individual curves for Types III, IV and V were quite erratic.

Also, the idea here is not to show that because of their large amount of mutual information, Type VI concepts are strictly easier than Type IV concepts, but that, in the long term, individuals might find strategies to simplify the Type VI categorization process. Hence, greater improvement in learning rates are expected for participants learning Type VI concepts than for those learning Type IV. It will turn out that Type VI is eventually easier than Type IV in the long run.

---

when some information is redundant between variables

The core hypothesis is that the computation of mutual information applies in the long run. Subjects would first use basic strategies such as finding a set of relevant features (or a combination of relevant features) to learn concepts. With repetition, subjects would then switch to more powerful strategies consisting in using mutual information to lower the complexity level of the task. Subjects would progressively discover relationships between dimensions that make the learning of concepts entailing high mutual information easier.

This paper relates two mixed-design experiments in three-dimensional Boolean concept learning, with one within-subject factor (repetition of learning of similar concept Types) and in which different concept Types were used as a between-subject factor (Experiments 1 and 2). In Experiment 1, subjects will learn repeatedly either a set of Type IV concept problems in a row, or a set of Type VI problems, using different stimuli for each problem. In Experiment 2, subjects will learn a set of identical Type IV concept problems in a row (or a set of Type VI concepts), using the exact same stimuli one problem after another, except that categories will be reversed for each problem. Experiment 3 is a total within-subject experiment in which subjects learned all 3D Boolean concepts once a week for one month.

Experiment 1

Each subject was required to learn either a set of Type IV concepts or a set of Type VI concepts during a one-hour period. The successive concepts will be called "Problems". The set of stimuli was different from one problem to another for a given subject. Only the Type of concept remained constant within subjects. Repeated measures on a single concept type were chosen so as to increase potential strategies based on mutual information. Subjects were asked to learn as many concepts as they could during the hour. No information was disclosed beforehand concerning the similarities between the successive problems.

Participants

The subjects were 20 Rutgers University students who received course credit in exchange for their participation. The subjects were randomly divided into two groups of equal size (Type IV vs Type VI).

Stimuli

In each problem, the stimuli varied along three separable binary dimensions (shape, color, and size). The two values for each dimension were chosen randomly from the following list (shape = triangle, square, or circle; color = blue, pink, red, or green; size = small or big). Each combination of values formed a single unified object stimulus (e.g., a small red square, a big blue circle). For each problem, the assignment of categories to the stimuli was chosen at random to conform to a Type VI or Type IV. By taking all combinations of values and all possible assignment of categories, 1152 Type IV concepts and 288 Type VI concepts could be potentially built. The stimuli were presented sequentially in blocks. In each block of $2^D = 2^3 = 8$ stimuli (D = number of dimensions), each stimulus appeared once in a random order, and the first stimulus of each block was different from the last of the previous block.

After a participant successfully learned the concept corresponding to the n^{th} problem, a new set of stimuli (using at least two different colors) was chosen for the $(n+1)^{th}$ problem (with exception of the classification Type which remained constant for the whole experiment for a given participant) and so on.

Procedure

The tasks were computer-driven. Participants learned to sort stimulus objects using two keys, with successful learning encouraged by means of a progress bar. The stimulus objects were presented one at a time in the upper part of the computer screen. After each response, feedback indicating a correct or incorrect classification was displayed at the bottom of the screen for two seconds. Subjects first learned the simplest concept in two dimensions in a short warm-up session (e.g., if black then positive, if white then negative).

The subjects scored one point in the progress bar for each correct response. A point was represented by an empty box that was filled each time a correct response was given. An incorrect response resulted in the loss of all points scored so far in the progress bar. This means that a 100% correct-classification criterion was adopted for each concept. Only such a criterion could guarantee that the subjects would learn all the objects equally well, regardless of their role in the concept. Due to the progress bar, subjects could not avoid responding to the most difficult stimuli and could not progress by classifying only the easy ones. To pace the learning process, each response had to be given in less than eight seconds. To make sure the subjects could also use the concept they had just learned, the number of points in the progress bar was equal to 4×2^D, that is four times the length of the training set of stimuli (as in the first experiment of Shepard et al., 1961). Subjects were therefore required to successively and correctly categorize all stimuli on four consecutive blocks of stimuli. The participants were rewarded with a digital image (animals, fractals, etc.) when they succeeded. They could continue (with a new problem) whenever they felt ready by clicking on a button. The subjects were stopped to make sure the experiment did not last more than one hour, including presentation and debriefing.

Instructions

The nature of the task was explained to the subjects: a set of stimuli would be presented repeatedly; subjects would have to learn their association with one of the two categories; the computer would give feedback with no tricks of any sort since the stimulus-response associations for one problem would be stable; their goal would be to complete the progress bar by giving successive correct responses; they would be informed as to when the next problem would appear since a next button would appear after they had completed the progress bar). Subjects were told that they were assigned 10 different problems for a maximum of one hour. Nothing was said about the heterogeneity of the categories in a given problem. Subjects were told that they would certainly find the task a bit difficult at first, but that they would certainly make progress from one problem to another.

Results

In short, on average, the Type VI concepts were learned in 8.1 blocks, as compared to 11.8 for the Type IV concepts. A more detailed analysis of the error rate per block for each problem is given in Figure 3 and Table 3 (the n^{th} problem corresponds to the n^{th} concept learned during the hour). The figure broadly indicates that the number of problems had an effect on learning for both concepts (the statistics for each curve are detailed below). The maximum number of problems completed in one hour was 15 (one subject only). Since the number of subjects decreased as the number of problems increased owing to time limitations, a limit of ten problems was chosen to keep a sample size above 10 subjects.

The overall comparison of the two learning curves reported in Figure 3 seems to support the hypothesis that, in the long run, subjects can benefit from the mutual information inherent in Type VI concepts. First, subjects averaged more errors during the first problem when learning Type VI ($M = 0.31$ –error rate–, $SD = 0.054$) than when learning Type IV ($M = 0.26$, $SD = 0.043$), before reaching the learning criterion ($t(18) = 2.53$, $\eta^2 = .262$, $p < .02$). This result has been unanimously confirmed in the earlier literature: Type VI is the hardest concept for subjects to learn. However, this trend changed rapidly in the present data: Type VI became easier than Type IV in all remaining problems. Considering the error rates for the first ten problems, Type VI ($M = 0.098$, $SD = 0.099$) turned out to be simpler than Type IV ($M = 0.180$, $SD = 0.094$); $t(172) = -5.7$, $\eta^2 = .162$, $p < .001$. This contradicts most models, which do not take relational information between dimensions into account. Note that this difference in error rates was even higher when the number of problems was used as a covariate; $F(1,172) = 42.9$, $\eta^2 = .199$, $p < .001$.

A better way to show that the mean error rate per block decreases more in the Type VI condition as the number of problems increases would be to compare the slopes of the regression lines that fit the two curves. Given that the slopes approached an exponential decay function, the natural logarithm of the number of errors was taken in order to minimize the sum of squares when fitting the error rates to the regression lines (the natural logarithm merely flattens the curves). A regression analysis was then run using the number of problems and two indicator variables that coded the types and the interaction between the problems and the types respectively. This technique –equivalent to an ANCOVA using problems as a covariate– allows one to test the null hypothesis of coincidence and the null hypothesis of parallelism of the two regression lines. Coincidence could hide either differences in parallelism or differences in intercepts. The result showed that the lines were not coincident ($F(2,171) = 24.2$, $\eta^2 = .081$, $p < .001$). This hypothesis being rejected, the hypothesis of parallelism could be tested. Again, the test allowed us to reject parallelism ($F(1,171) = 4.6$, $\eta^2 = .015$, $p < .05$). This means that the slope was significantly greater downward for Type VI than for Type IV.

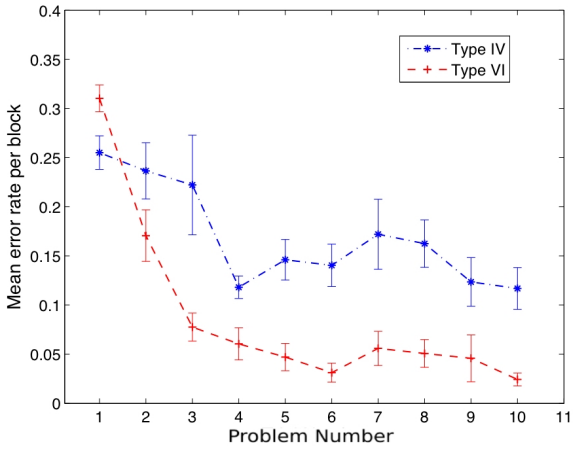


Figure 3. Mean error rate per block for subjects to reach criterion for the first 10 problems. *Note.* Error bars indicate +/- one standard error.

Experiment 2

In experiment 1, mnemonic strategies using numerical relationships might have been used by subjects. For instance, subjects might have taken a positive stimulus as a reference and computed the number of common features between this stimulus and another stimulus to determine the membership of the latter. For instance, let us consider the 010 stimulus at the lower left of the front face of the Type IV and the Type VI cubes in Fig 1. In the Type VI concept, one feature in

Table 3
Mean error rate per block in 10 successive problems of the same Type (Type IV or Type VI) in Experiment 1

	Problem Number									
	1	2	3	4	5	6	7	8	9	10
TYPE IV										
Error rate	.255	.237	.222	.118	.146	.141	.172	.163	.124	.117
SD(Error rate)	.054	.091	.160	.035	.062	.057	.094	.064	.066	.052
TYPE VI										
Error rate	.310	.171	.078	.061	.047	.031	.056	.051	.046	.024
SD(Error rate)	.043	.083	.045	.052	.044	.030	.055	.042	.067	.016

Note. The mean error rate is the mean number of incorrect responses per block computed for a given problem, divided by the number of examples in a block (8).

common with the 010 stimulus signals a positive stimulus; in concept IV, two features in common with 010 signals a positive stimulus. Such numeric information can be used and can lead to very short rules in Type IV and VI, but most of the time, it would lead to unproductive strategies when learning Boolean concepts (For instance, computing the number of common features in a Type I concept would be counterproductive). Shepard et al. had independent judges rate the amount of unnecessary complexity in the rules verbalized by the subjects in their first experiment which tended to prove that, even if the greatest reduction of complexity was observed for Type VI across problems, the subjects generally did not used such numerical rules since the mean rating of unnecessary complexity in the rules was quite high for Types IV and VI and higher in Type VI. In their second experiment, none of the subjects discovered the numerical rule in Type VI. Still, the use of numeric strategies by certain subjects might surrogate the rules and their complexity defined by computational models. This experiment aims at enhancing and suppressing such numerical facilitation. To anticipate, in comparison with the compound condition used in Experiment 1, a spatial condition enhancing the numerical facilitation and a narrative condition suppressing the numerical facilitation will be assigned to subjects (this is detailed below).

Also, the Type VI concept might have been found easier by subjects across problems, because there were fewer variations in this kind of problems, in part due to the limitation of the number of values in the size dimension. On one hand, the focus of attention on the size dimension could have favored the Type IV problem. For instance, using size as a relevant dimension straight away in a Type IV problem guaranteed 6 correct responses, because there are always three positive examples for a particular size and three negative for the other size value for any rotation. Therefore, perhaps the learner only needed to memorize one exception per size value. On the contrary, the size dimension alone would only give four correct responses in a Type VI problem if the subject had tried to classify the stimuli according to their size. This possibility might have interfered with the results but could not explain per se the greater ease of learning in Type VI problems.

But on the other hand, in type VI problems, two stimuli of the same size and shape (but different in color) were necessarily in different categories. Subjects could then use some analogies between problems (for instance, for the small objects of a given shape, subjects only had to find which colors matched the categories). Thanks to the effect of mutual information, the subjects could perfectly determine the structure of the entire category from the feedback obtained from the first stimulus of a problem (from the second problem on). On the contrary, in Type IV concepts, the presentation and feedback for the first stimulus was not sufficiently informative to predict all the next category responses. At least three or four stimuli were necessary to induce the correct rule using a rule analog to the one used in the previous problem. The same problem arose in Experiment 1 by Shepard et al., because the five consecutive problems were associated with five set of different stimuli, but only two values were used per dimension. Using more dimensions and more dimension values would not solve the problem as the same strategies could be applied by the subjects once the irrelevant dimensions are identified. To address this issue, Experiment 2 was designed to remove this factor by using identical stimuli across problems.

A last problem in Experiment 1 might be due to the use of a warm-up concept, which might have made the subjects search for rules at the beginning of the first problem, which is more incompatible with Type VI. This could explain its greater difficulty during the first problem.

Participants

The subjects were 84 University of Franche-Comté students who received course credit in exchange for their par-

ticipation. The subjects were randomly assigned to 6 groups of 14 subjects.

Stimuli

In this experiment, the learning of Type IV and Type VI classifications was investigated in three different stimulus conditions. Stimulus conditions varied according to the degree of relationship that could be defined between the stimuli (cf. Fig. 4). In the first condition (Spatial condition), the subjects were shown three similar objects (three black balls) placed above or under a line. Boolean values across dimensions were up versus down (instead of the black/white opposition shown in Table 1). This condition highlighted the correlations between dimensions, as noted in section "Simplicity of Relational Concepts: XOR and Type VI" and in Table 1. Mutual information was enhanced by the spatial position of the balls around the line. In this case, subjects could devise numeric or spatial strategies specially efficient in Type VI. For instance, in the Type VI concept, the subjects could notice that the 2^{nd} and the 3^{rd} balls were in different locations whenever the 1^{st} ball was under the line, whereas the balls were in the same location (all above the line or all under) whenever the 1^{st} ball was above the line. This also created a situation in which only the stimulus was positive whenever the number of balls above the line was odd in Type VI concepts. The numeric and spatial features could also be used, to a lesser extent in Type IV concepts.

In the second condition (Narrative condition), the stimuli (cf. Fig. 4) were built in order to allow a narration of the different events depicted in the stimuli. Totally different dimensions were used to reduce commonalities between dimension values. The first dimension represented two different seasons (summer vs winter). Several features were associated with each season (green grass, blue sky and sun *vs* snow, gray sky and clouds). The second dimension represented what could be offered by a boy to the girl protagonist (a cake vs a bouquet of flowers). The third dimension was the girl protagonist (Mary, in a white dress vs Lucy, in a pink dress). In this condition, it was possible for subjects to elaborate little scenarios to devise mnemonic strategies (such as: Mary likes flowers in summer and cakes in winter, etc.). Subjects were given the list of the dimensions which were subject to variations at the beginning of the experiment. This condition was expected to suppress some numerical strategies such as those that were more obvious in the Spatial condition (although computing the number of commonalities between successive pictures was still possible). In the third condition (Compound condition), stimuli similar to those used in Experiment 1 were built, in order to allow some comparison.

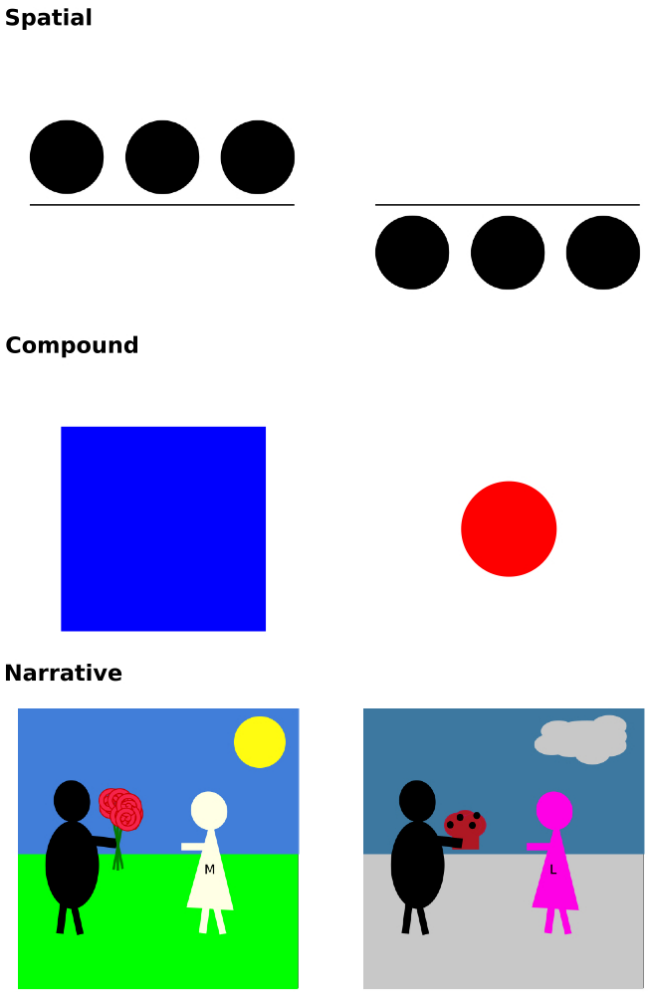


Figure 4. Sample of stimuli used in Exp 2. *Note.* In each row, the two stimuli are opposed on each of their varying dimensions.

Procedure

Each subject was given a series of 10 similar concepts. The experiment was designed to last one hour, including presentation and debriefing. The subjects were not required to complete the 10 problems, but were asked to complete the maximum number of problems in the time that was allocated to the experiment. Half of the subjects was assigned to a series of Type VI concepts, the other half to Type IV. Contrary to experiment 1, there were no variations in the stimulus features between problems for a given subject. The same stimuli were given problem after problem. The only variation between problems was the stimulus-category mapping. Category membership was simply reversed from one problem to the other. Hence, whenever the problem number was even,

the positive examples of the previous problem became negative (and vice-versa for the negative examples). The subjects simply encountered the same category structure for every two problems. The goal was to measure the ability of the subjects to reverse the category membership across problems in a single concept Type.

In the three conditions, for each subject, a set of positive examples was randomly chosen for the first problem to conform to a Type IV or Type VI structure. The positive examples were then memorized by the program and successively reversed from one problem to the next. In the compound condition, for each subject, the stimulus features were randomly drawn (from a set of features equivalent to the one used in Experiment 1) at the beginning of the experiment and were then maintained constant throughout the task.

There was no warm-up session in this experiment, to avoid orienting subjects to search for simple rules. Otherwise, the procedure was identical to that of Experiment 1 (feedback was given for 2 seconds, a progress bar indicated the number of successive correct responses, each stimulus appeared once in a block, and subjects were required to correctly categorize all stimuli on four consecutive blocks).

Results

Figure 5 shows the learning curves in the Spatial, Compound and Narrative conditions (frames *A*, *B* and *C*) when the error rate per block was measured. When the first problem was analyzed separately, a Condition (Spatial, Compound, Narrative) by Type (IV vs VI) ANOVA found only a significant effect of Condition, $F(2,78) = 5.93$, $p = .004$, $\eta_p^2 = .13$, in which means equal to .28 ($sd = .07$), .32 ($sd = .05$) and .34 ($sd = .07$) respectively. During the first problem, Type VI was found more difficult than Type IV ($t(26) = 2.54$, $p < .017$, $\eta^2 = .20$) only in the Compound condition. This result matches our observation in the first experiment, in which stimuli were also compound. In the other two other conditions, no difference between Type IV and Type VI was observed during the first problem.

Because the overall improvement over problems was not of crucial interest in our study, the data were collapsed across the ten problems (cf. Fig. 5.D). Then, another Condition (Spatial, Compound, Narrative) by Type (IV vs VI) ANOVA was run. This time, a main effect of Type was observed (Type VI being easier, $F(1,631) = 25.95$, $p < .001$, $\eta_p^2 = .04$), a main effect of Condition (Spatial < Compound < Narrative, $F(1,631) = 22.9$, $p < .001$, $\eta_p^2 = .07$; the post hoc pairwise comparisons between the Spatial, Compound and Narrative conditions were all significant using the Bonferroni adjustment), as well as a significant interaction ($F(1,631) = 4.35$, $p < .05$, $\eta_p^2 = .01$). The interaction indicated a greater discrepancy between Type VI and Type IV in the Narrative condition.

Table 4 shows the means and standard deviations of these collapsed results. Overall, the odds ratios (each ratio was computed between the odds of error rate in Type IV and the odds of error rate in Type VI in a given condition) were 1.3, 1.2, and 1.8 respectively, in the Spatial, Compound, and Narrative conditions. For instance, the proportion of errors was 1.3 times larger ([.127/.873]/[.103/.897] = 1.3) in Type IV than in Type VI in the Spatial condition. Overall, the Spatial condition and the Narrative condition were beneficial and detrimental to learning respectively. When taking a closer look at the Spatial condition compared to the Compound condition, there is a significant difference between the Spatial and Compound conditions within each rule type ($t(255) = 2.73$, $p < .007$, for Type VI; $t(255) = 2.83$, $p < .005$, for Type IV). These differences appear to be of the same magnitude because there is no statistical interaction between Type and Condition, when the analysis is restricted to the Spatial and Compound treatments. Because the Spatial condition facilitated learning of both Types, numerical biases can then be considered orthogonal to mutual information. On the contrary, the Narrative condition (compared to the Compound condition) hindered learning of Types IV ($t(162) = 4.02$, $p < .001$) more than learning of Type VI ($t(185) = .45$, *NS*). Hence, the Narrative condition increased the disparity between the two Types, which is confirmed by an interaction effet ($F(1,347) = 6.22$, $p = .013$, $\eta_p^2 = .02$). In other words, the easiness of Type VI in the Narrative condition might be due to the use of pure (i.e., non numerical) relational information. To conclude, mutual information was beneficial for Type VI rules across a wide range of stimulus types.

Analysis of previous data (Experiment 3)

The following is an analysis of an experiment which differed in several aspects from the two preceding ones. This experiment was carried out by Mathy (2002). Fourteen subjects were asked to learn the 13 three-dimensional Boolean concepts once a week over a period of four weeks (for a total of $4 \times 13 = 52$ concepts). Each week, the concepts were given in random order. The first concept of the $(n+1)^{th}$ week was not identical to the last concept of the n^{th} week. The experimental setting was similar to the one used in Experiment 1, except that the learning criterion was based on two consecutive blocks of successive correct responses instead of four. Stimuli features were randomly chosen for each of the 52 concepts which were learned. The stimuli were different from the ones used in Experiment 1: stimuli features were oval, triangular, or square shapes; shapes were shown in pink, green, blue or red colors; the last dimension was an circle vs diamond frame around the central shape. Basically, the frame dimension replaced the size dimension used in Ex-

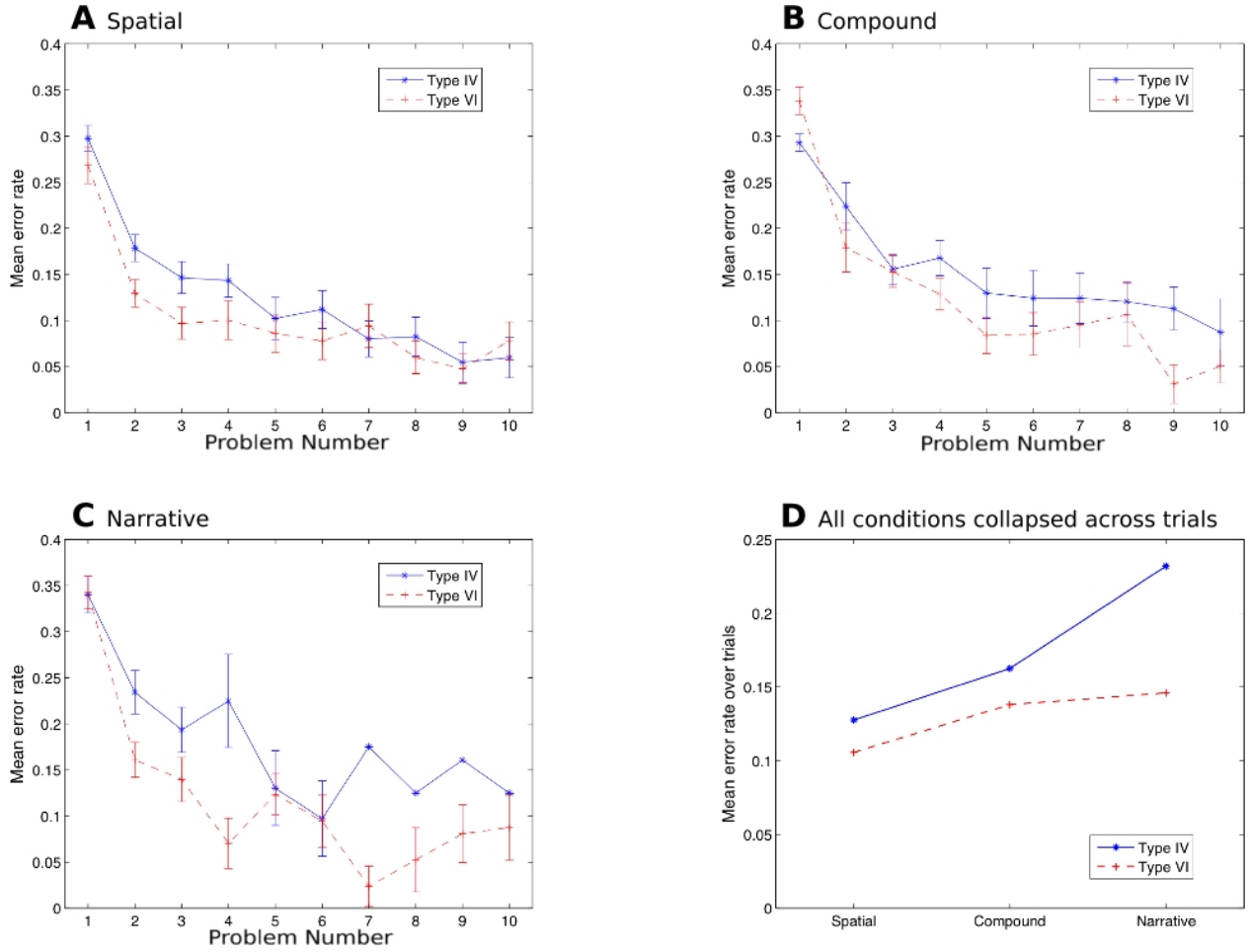


Figure 5. Mean error rate over problems for Type IV and Type VI concepts, in three different stimulus conditions. *Note.* Error bars indicate +/- one standard error.

Table 4
Mean error rate (and SDs) in 10 successive problems of the same Type (Type IV or Type VI) in three conditions (Spatial, Compound, Narrative), measured in Experiment 2.

	Spatial	Compound	Narrative
Type IV	0.127	0.163	0.232
	(0.099)	(0.097	(0.112)
Type VI	0.103	0.138	0.146
	(0.094)	(0.112)	(0.122)
Odds ratio	1.3	1.2	1.8

Note. The error rate was computed across all the problems performed by a subject. The odds ratio for the Spatial condition is: [.127/(1-.127)]/[.103/(1-.103)] = 1.3; The Total column cannot be computed from the mean of the three precedent columns without weight adjustment, because the number of blocks or problems were different in the conditions

periment 1. This experiment differs from the two preceding ones in that there were no successive similar concept Types, so the subjects could not devise strategies which applied to a single Type of concept throughout the experiment.

Only a subset of the concepts of interest are analyzed next. The data was pooled according to whether the concepts were characterized by null information (Types I and II) or positive mutual information (Types IV, V and VI). The number of blocks was subjected to a logarithmic transformation to limit the positive skewness of the distributions. The results were then plotted as if an ANCOVA was about to be computed, with week as a covariate. The reason is that before conducting an ANCOVA, the assumption of equal slopes needs to be tested first. The ANCOVA effectively assumes that there is no interaction between the covariate and treatments. The results plotted in Figure 6 indicate that the covariate (week/problem number) by the treatment (null vs positive mutual information) interaction is significant ($F(1,172) = 7.2$, $p = .008$), meaning that the regression slopes are not equal. The study of the effect of negative mutual information goes beyond this study, but it can be noted that Type III, characterized by negative mutual information (hence entailing some causality between features) showed a slope in between. In conclusion, the subject's performance in categorization tasks entailing mutual information decreased with time and repetition, even when learning was disrupted by intercalated concepts of a different nature.

It could be argued that in this experiment, progression might simply be correlated to concept difficulty, with a greater progression margin for complex Types and a floor effect for others (effectively, the partial correlation between repetition and mutual information controlling for concept complexity was not significant in this data). However, when the previous experiments are taken into account, this experiment gives some extra information on the putative effect of mutual information on learning.

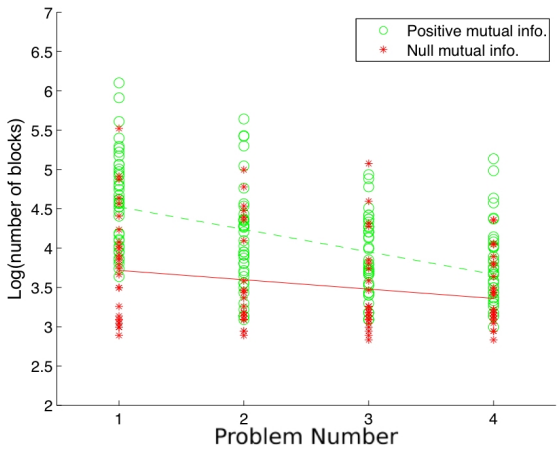


Figure 6. Logarithm of the number of blocks to reach the criterion against treatment (Null vs positive mutual information) and covariate (problem number, or week). *Note.* Concepts entailing null mutual information are Types I and II. Concepts entailing positive mutual information are Types IV, V, and VI.

Discussion

The present results suggest that relational information in concepts is relevant to learning a sequence of successive problems of the same type. Our results simply show a correlation between performance in classification and the presence of relational information between input variables in Boolean concepts (Experiment 3). In particular, the two first experiments showed that because Type VI problems entail a large amount of mutual information, Type VI can be easier to learn than Type IV problems, in the long run. In the present experiments, one may have noted that the number of stimuli (8) and the number of dimensions (3) were constant. This explains why increasing the amount of information that can be transmitted between variables in some concepts enabled greater performance in our study, whereas in other studies on absolute judgments (Miller, 1956), the amount of information that can be transmitted by subjects increases initially as the number of stimuli is increased, but gradually levels off because of the limitation of the channel capacity in subjects.

The present study confirms other research which has shown that within-category correlations can be learned during classification tasks, even incidentally (Giguère, Lacroix, & Larochelle, 2007). Relational information can be measured easily by mutual information, which is a very convenient non-metric tool for computing the amount of information shared by variables in multivariate distributions. Using repeated measures, the present study probes into the long-term effect of mutual information, and points out that learning several classifications of the same Type in succession has a substantial impact on how concepts are learned. By comparing Type VI concepts (entailing a maximal amount of positive mutual information) with Type IV concepts (entailing a minimal amount of positive mutual information), the present results show that Type VI concepts gradually become more learnable than Type IV ones. For instance, from the second problem on, the error rate for Type VI was lower than for Type IV in Experiments 1 and 2. These results suggest that the categorization models presented in the introduction needs to be refined to take the effect of relational information into account. This work emphasizes the peculiar status of Type VI concepts, which has also been noted in different research on inductive biases and cultural evolution (Griffiths, Christian, & Kalish, 2008; Griffiths, Kalish, & Lewandowsky, 2008; the authors have shown that in transmission chains,

in which individuals pass information to others, people converge more often than expected towards Type VI).

A possible interpretation of the greater difficulty with Type VI concepts on the first categorization task is that subjects could have spent time looking for rules or regularities and resisted having to memorize all examples. This might apply to any experiments in which Type VI was found to be more difficult. Effectively, subjects might use natural rule-like strategy, which, in general, is more compatible with any other Type, and which could explain greater performance for Type IV during the first problem in our experiments. However, this alone would not explain why performance on Type VI concepts increases on the next problems. If concepts were learned via a process akin to overt memorization on the next problems, learning of Type VI would remain more difficult than Type IV since Type IV could still benefit from more abstraction. As previously shown in the literature on concept learning, there would be no difference between classification Types if nothing but rote memorization were used by subjects. The greater simplicity of Type VI therefore calls for another explanation. Also, there is no reason why mutual information could not be applied after a couple of blocks during the first exposure, rather than after a couple of successive problems. Therefore, a more rapid learning of Type VI during the first trial (observed in Experiment 2) is not incompatible with the explanation based on mutual information suggested here.

Relations are of great interest in processes such as reasoning with polyadic predicates like "x is taller than y" (Goodwin & Johnson-Laird, 2005), making higher order inferences such as "x loves y more than z does" (Goodwin & Johnson-Laird, 2006), analogie formation (Gentner, 2006), and also in categorization (Gentner & Kurtz, 2005). The present study addresses an important issue in categorization modeling: Most models focus on single metrics. So far, no model in psychology has been designed to predict the use of relational complexity in classification learning in connection with any of the classical rule-based or similarity-based models. Halford and colleagues (Halford, Wilson, & Phillips, 1998; Halford, Cowan, & Andrews, 2007) also suggest the use of relational metrics, although the relational complexity defined by these authors is a bit different and should rather be called "interactional" because it refers to the minimal dimensionality to which a representation can be reduced without losing the information necessary for a solution (for instance, the XOR corresponds to a first-order interaction, because the two input values are necessary for the categorization process of all stimuli). This notion applies here: the Type VI concept can be described as a second-order interaction, noticeable in the symmetries in the decision tree in Figure 1. Hence, not only does mutual information connect with relational complexity, Bayesian networks, and decision processes as explained in the introduction, but also easily with interactional complexity.

This allows us to propose the idea of investigating, combining and integrating different metrics in categorization modeling. Some attempts have already been made to produce hybrid models using both rule-based metrics and similarity-based metrics (Anderson & Betz, 2001; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Nosofsky, Palmeri, & McKinley, 1994; Smith & Sloman, 1994). The present study suggests that more effort should be applied to including relational metrics. A key point in predicting future results is the applicability of the different strategies: some strategies might be used quickly (e.g., attempting to use single features), whereas others might require more time (i.e, finding relations, correlations, or causal connections). Also, potential transitions between strategies need to be examined in greater detail in the future.

A much more delicate point concerns the differential compatibility of models with relational information. The present results tend to argue in favor of categorization models with metrics based on abstraction and compression. An information reduction process could effectively explain how multiple strategies might occur and result in transfer effects. For instance, a minimal disjunctive formula such as $xy + x'y' = 1$ (i.e., IF (x and y) OR (x' and y') THEN the example is positive), which entails some mutual information, can easily be reduced to $(X = Y) = 1$ (i.e., if the input dimensions are the same, then the example is positive). The present results do not provide a definite answer, but it is worth mentioning that relational information cannot be easily accounted for in exemplar-based models, as xy and $x'y'$ are maximally dissimilar. This study calls for a more complete analysis of within-type transfer on a much larger class of concepts such as the Feldman's family set (Feldman, 2003). An extended data set would provide better evidence for a subject's reliance on mutual information.

References

Anderson, J. R., & Betz, J. (2001). A hybrid moel of categorization. *Psychonomic Bulletin & Review*, *8*(4), 629–647.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442-481.

Attneave, F. (1959). *Applications of information theory to psychology: a summary of basic concepts, methods, and results*. New York, NY: Holt, Rinehart, and Winston.

Bourne, L. E. J. (1970). Knowing and using concepts. *Psychological Review*, *77*, 546-556.

Bradmetz, J., & Mathy, F. (2008). Response times seen as decompression times in Boolean concept use. *Psychological Research*, *72*, 211-234.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New York: Wiley.

Bryant, R. (1986). Graph-based algorithms for Boolean function manipulation. *IEEE Transactions on Compilers, C-35*, *8*.

Coombs, C. H., Dawes, R. M., & Tversky, A. (Eds.). (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. New York, NY: John Wiley and Sons.

Estes, W. K. (1994). *Classification and cognition*. New York, NY: Oxford University Press.

Fass, D. (2006). *Human sensitivity to mutual information*. Unpublished doctoral dissertation, Rutgers University.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, *47*, 75-89.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology*, *50*, 339–368.

Garner, W. (1962). *Uncertainty and structure as psychological concepts*. New York: John Wiley and Sons.

Gentner, D. (2006). Relations, objects, and the composition of analogies. *Cognitive Science*, *30*, 609-642.

Gentner, D., & Kurtz, K. J. (2005). Categorization inside and outside the laboratory. In D. L. Medin, W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), (p. 151-175). Washington, DC: APA.

Giguère, G., Lacroix, G. L., & Larochelle, S. (2007). Learning the correlational structure of stimuli in a one-attribute classification task. *European Journal of Cognitive Psychology*, *19*, 457-469.

Glymour, C. N. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Massachusetts, MA: MIT Press.

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

Goodwin, G. P., & Johnson-Laird, P. N. (2006). Reasoning about the relations between relations. *The Quaterly Journal of Experimental Psychology*, *59*, 1047-1069.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*, 68 — 107.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical transactions of the royal society*, *363*, 3504-3514.

Halford, G., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: a new hypothesis. *Trends in Cognitive Sciences*, *11*, 236-242.

Halford, G., Wilson, W., & Phillips, W. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803-831.

Houtsma, A. J. M. (1983). Estimation of mutual information from limited experimental data. *Journal of the Acoustical Society of America*, *74*, 1626-1629.

Hovland, C. (1966). A communication analysis of concept learning. *Psychological Review*, *59*, 461-472.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Lafond, D., Lacouture, Y., & Mineau, G. (2007). Complexity minimization in rule-based category learning: Revising the catalog of boolean concepts and evidence for non-minimal rules. *Journal of Mathematical Psychology*, *51*, 57-74.

Lehrl, F., & Fischer, B. (1990). A basic information psychological parameter (bip) for the reconstruction of concepts of intelligence. *European Journal of Personality*, *4*, 259-286.

Luce, R. (1963). Handbook of mathematical psychology. In R. Luce, R. Bush, & E. Galanter (Eds.), (p. 103-190). New York: Wiley.

Luce, R. (2003). Whatever happened to information theory in psychology. *Review of general psychology*, *7*, 183-188.

Mathy, F. (2002). *L'apprenabilité des concepts évaluée au moyen d'un modèle multi-agent de la complexité des communications en mémoire de travail*. Unpublished doctoral dissertation, Unpublished doctoral dissertation, Université de Reims, France.

Mathy, F., & Bradmetz, J. (2004). A theory of the graceful complexification of concepts and their learnability. *Current Psychology of Cognition*, *22*, 41-82.

Medin, D. L., & Schaffer, M. (1978). A context theory of classification learning. *Psychological Review*, *85*, 207-238.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*, 343-352.

Mori, S. (1998). Effects of stimulus information and number of stimuli on sequential dependencies in absolute identification. *Canadian Journal of Experimental Psychology*, *52*, 72-83.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rules-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352-369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53-79.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York, NY: Cambridge University Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423, 623-656.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 13, whole No. 517.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3-22.

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neurosci. Biobehav. Rev.*, *32*, 249-264.

Smith, E. E., & Sloman, S. A. (1994). Similarity- vs. rule-based categorization. *Memory & Cognition*, *22*(4), 377–386.

Vigo, R. (2006). A note on the complexity of boolean concepts. *Journal of Mathematical Psychology*, *50*, 501-510.

APPENDIX

Entropy, joint entropy, and conditional entropy

Before describing mutual information, some of the main concepts behind information theory should be examined. Information theory determines the quantity of data a set of messages contains by computing basic probabilities. This quantity called entropy (or information content) corresponds to the amount of uncertainty one has about a set of messages. For example, someone knows whether a card is black or red. Because the probability of guessing the color is .5, the communication of a single piece of information that is communicated (e.g., black) is $-log_2(.5) = 1$, also called the surprisal. Therefore, a single message (black or red) corresponds to 1 bit of information. The uncertainty one has about a set of possible messages (e.g., black or red) is called the entropy H. It is determined by computing the expected value of the surprisal of all possible pieces of information a message might contain. Given:

$$H(X) = -\sum_i p_i log_2(p_i) \tag{5}$$

we have: $H(color) = -p(red)log_2(p(red)) - p(black)log_2(p(black)) = -(\frac{1}{2})(-1) - (\frac{1}{2})(-1) = .5 + .5 = 1$. H corresponds to the number of binary questions one would need to ask to retrieve the content of a message. Here the question would simply be: Is the card red or black?

Let us imagine someone knows whether a card is spades, hearts, diamonds or clubs. In this case, $H(suit) = -\sum .25 \times log_2(.25) = 2$. Effectively, two binary questions are necessary to discover the suit of a card (Is the card red or black? Then, if the color is black: Is the card Spades or Clubs?). This also means that for coding 4 symbols, only two binary variables are necessary (00 = hearts, 01 = diamonds, 10 = clubs, and 11 = spades).

Take the example of the XOR structure shown in Fig. 2. Each column is a variable that can take 0 or 1 values (the variable can be used to send messages composed of a single 0 or 1). If these columns are renamed X, Y, and Z:

$$\textbf{XOR} = \begin{pmatrix} X & Y & Z \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

The entropy H of each of these variables is 1 bit because 1 bit of information is needed to store or communicate one of the two equally probable values (0 or 1) that can be taken by the variable. Given:

$$H(X) = -\sum_{i-1}^{m} p_i log_2(p_i) = \tag{6}$$

we have: $H = -p(0)log_2(p(0)) - p(1)log_2(p(1)) = -(\frac{1}{2})(-1) - (\frac{1}{2})(-1) = .5 + .5 = 1$.

It is also possible to compute the joint entropy or the conditional entropy of variables. The joint entropy is simply the entropy of the set of messages that can created be using several variables. Again, using two binary variables, it is possible to form 4 different messages $\{00, 01, 10, 11\}$. In this case, where variables are independent, the joint entropy (the entropy of the conjunction of variables) is simply the sum of their individual entropies: $H(X,Y) = H(X) + H(Y) = 2$, the comma symbol between X and Y indicating the conjunction between X and Y. The conditional entropy $H(X/Y)$ (the slash symbol indicating the conditional statement "knowing") is a measure of the quantity of information in one variable holding a second constant. For instance $H(X/Y) = 1$, because holding Y constant leaves 1 bit of uncertainty. For instance, for $Y = 1$, X is either 0 or 1; idem for $Y = 0$.

Mutual information

Mutual information is a measure of the quantity of information one can obtain on a given set of variables by observing another variable. The mutual information between two variables is:

$$I(X;Y) = H(X) - H(X|Y) \tag{7}$$

With :

$$H(X|Y) = H(X,Y) - H(Y) \tag{8}$$

The variables are separated by semicolons in the formula to avoid confusion with conjunctions. For any pair of variables in the XOR truth table (e.g., X and Y), we get null mutual information. For instance: $I(X;Y) = H(X) - H(X|Y) = H(X) - (H(X,Y) - H(Y)) = 1 - (2 - 1) = 0$, meaning that these two variables are independent. Hence, taken by pairs, X, Y and Z are independent.

Even if it looks much more complicated, the computation of mutual information is easily extendable to an arbitrary

number of dimensions using alternating plus and minus signs over all subsets of variables. For three variables:

$$I(X;Y;Z) = H(X,Y) + H(X,Z) + H(Y,Z) + \\ - H(X) - H(Y) - H(Z) - H(X,Y,Z) \qquad (9)$$

Computed for the three variables in the XOR truth table, we have:
$I(X;Y;Z) = 2+2+2-1-1-1-2 = 1$, which corresponds to the maximal amount of mutual information with three Boolean variables. As explained throughout the article, this means that it is possible to know the value of a given variable given the relationship between the two others (or vice-versa).