Linguistic barriers to logical reasoning: a new perspective on Aristotelian syllogisms[1]

Andreas HAIDA — *ELSC, The Hebrew University of Jerusalem*
Luka CRNIČ — *LLCC, The Hebrew University of Jerusalem*
Yosef GRODZINSKY — *ELSC, The Hebrew University of Jerusalem*

Abstract. Experimental studies investigating logical reasoning performance show very high error rates of up to 80% and more. Previous research identified scalar inferences of the sentences of logical arguments as a major error source. We present new analytical tools to quantify the impact of scalar inferences on syllogistic reasoning. Our proposal builds on a new classification of Aristotelian syllogisms and a closely linked classification of reasoning behaviors/strategies. We argue that the variation in error rates across syllogistic reasoning tasks is in part due to individual variation: reasoners follow different reasoning strategies and these strategies play out differently for syllogisms of different classes.

Keywords: syllogisms, reasoning errors, individual variation, scalar inferences.

1. Introduction

Our paper investigates the impact of so-called *scalar inferences* on logical reasoning performance. From almost its outset, the study of the psychology of logical reasoning aimed at identifying common inferences that lead to divergence from logically valid reasoning (Sells, 1936; Wilkins, 1928; Woodworth and Sells, 1935). A long-recognized example of such an inference is the scalar inference (SI) from the truth of an existential sentence to the falsity of its universal counterpart, represented by the scheme in (1) (Begg and Harris, 1982; Newstead and Griggs, 1983; Rips, 1994).[2]

(1) some Ms are Ks $\overset{\text{SI}}{\rightsquigarrow}$ not all Ms are Ks

To see how commonly the SI in (1) seems to be drawn in logical reasoning tasks, consider the argument in (2), from the premise (I) to the putative conclusion (O).[3] We will be looking at the

[1]We would like to thank the participants of the "Semantics and Natural Logic" special session of *Sinn und Bedeutung 22* for their comments and suggestions. Versions of this work have been presented at the LINGUAE Seminar at Institut Jean Nicod, Paris and the linguistics seminar at the University of Wisconsin at Milwaukee. We wish to thank the audiences of these events for their constructive remarks. This work was supported by grant 1926/14 from the Israel Science Foundation (Crnič), grant 2093/16 of the Israel Science Foundation (Grodzinsky), and a post-doctoral grant from the Edmond and Lily Safra Center for Brain Sciences (Haida).

[2]An argument that the inference in (1) is not a logical entailment comes from the fact that the negation of an existential sentence is false if its universal counterpart is true. This means that either the indefinite determiner *some* is not a logical constant of English or the SI of an existential sentence is not a logical entailment. In either case, the inference in (1) is not logically valid. We follow the standard assumption due to Grice (1975) that *some* is a logical constant, and hence that the inference in (1) is not a logical entailment. See footnote 8 for further discussion.

[3]Here and below, we use the letters I and O to designate existential sentences by their traditional names from Aristotelian-scholastic logic (see §2 for a compilation of relevant terminology). The difference between I- and O-sentences is that the predicate in the nuclear scope of the indefinite determiner is negated in O-sentences and

 In: Uli Sauerland and Stephanie Solt (eds.), *Proceedings of Sinn und Bedeutung 22*, vol. 1, ZASPiL 60, pp. 453–468. ZAS, Berlin.

impact of the SI of the I-sentence, i.e., at the impact of (1). Take note that the I-sentence functions as the *premise* of the argument, since this will play an important role for our discussion.

(2)
$$\frac{\text{(I)}\quad \text{Some Ms are Ks}}{\text{(O)}\quad \text{Some Ms are not Ks}}$$
accepted by 94% of all subjects

Newstead and Griggs (1983), henceforth N&G, report that, when asked to decide whether the O-sentence "logically follows" from the I-sentence, 94% of all subjects gave a positive response.[4] However, inferences from I-sentences to O-sentences are not logically valid: in Aristotelian logic (AL) as well as in predicate logic (PL) (and all other logics that we are aware of) existential sentences are logically compatible with their universal counterparts.[5,6] Thus, since the reasoning task targeted logical inferences,[7] 94% of all subjects erred in their judgment. A possible explanation for this high error rate is that the vast majority of subjects not only considered the logical entailments of the I-premise but also its SI.[8] Importantly, the conjunction of the I-premise and its SI logically entails the O-conclusion.[9] This observation suggests that the errors observed in the I-to-O inference task are due to the SI of the I-premise (as already concluded by N&G). Furthermore, the magnitude of the error rate suggests that almost all reasoners computed (and took it into account) the SI of the I-premise.

Next, we are looking at the same reasoners and the same I-sentence, but this time when it functions as the *conclusion* of an argument. N&G observe that the argument in (3), from the premise (A) to the putative I-conclusion, was judged logically valid by 73% of all subjects.[10] That is, as indicated only 27% of all subjects rejected the validity of the A-to-I inference.

---

not negated in I-sentences.

[4]In the experiment instructions, subjects were informed that alphabetic letters, in our example *M* and *K*, stand for classes of things. In a follow-up experiment, N&G found that replacing the letters with concrete nouns such as *artist* or *bee-keeper* does not lead to significantly different results.

[5]The language of AL is a proper fragment of the language of PL. Semantically, AL differs from PL in that universal sentences entail their existential counterparts. Thus, in the former *all Ms are Ks* entails *some Ms are Ks* and *no Ms are Ks* entails *some Ms are not Ks*. Natural language quantifiers are Aristotelian in the sense that they entail (or presuppose) that the extension of their restriction is non-empty. Still, there might be reasoners who employ PL in logical reasoning tasks.

[6]In AL, the logical compatibility of existential sentences with their universal counterparts implies that the inference in (1) is not an entailment (cf. footnote 2).

[7]Maybe problematically, N&G's experiment instructions do not spell out what it means for a sentence (form) to logically follow from another one. Still, if the determiner *some* is a logical constant of English (see footnote 2) N&G's result can be taken to show that only few subjects assigned the I-sentence its logical meaning, $\exists x(\mathrm{M}x \wedge \mathrm{K}x)$.

[8]According to the grammatical view of SIs, the I-premise can entail the O-conclusion, namely if and only if the string *some Ms are Ks* is parsed with a covert exhaustification operator *exh* (Chierchia et al., 2012). This means that the grammatical view also holds that the indefinite determiner *some* is a logical constant of English and that its truth-conditional content does not bring about the inference in (1) all by itself. Hence, since I-sentences can be parsed without *exh*, the inference from I- to O-sentences is not a *logical* entailment on the grammatical view either.

[9]The SI of the O-conclusion, viz. that its stronger universal counterpart *all Ms are not Ks* ($\equiv$ *no Ms are Ks*) is false, is entailed by the I-premise. Thus, the judgment whether the O-sentence logically follows from the I-sentence is not affected by the SI of the O-sentence.

[10]The letter A is the traditional name of universal affirmatives, i.e., sentences of the form *all Ms are Ks*. Universal negatives, i.e., sentences of the form *no Ms are Ks* will be designated by the letter E (see also §2).

(3)
$$\frac{\text{(A)}\quad \text{All Ms are Ks}}{\text{(I)}\quad \text{Some Ms are Ks}}$$
rejected by 27% of all subjects

In AL, A-to-I inferences are logically valid. However, the I-conclusion can only be drawn if its SI is *not* computed. Thus, the rejection rate of the A-to-I inference suggests that only a minority of reasoners computed the SI of the I-conclusion.[11]

These observations raise the question of why the SI of an I-premise is computed more frequently than the SI of an I-conclusion, i.e., $> 90\%$ vs $< 30\%$ of all times.[12] Our answer will be based on the consideration that the locus or loci of SI computation characterize different types of reasoners, viz. the four types in Table 1, where the rows $\pm$strong mark whether or not the SI of a premise is computed and conjunctively added to its literal meaning, and likewise for the columns and the conclusion. (Henceforth, instead of saying that the SI of a sentence is computed and conjunctively added to its literal meaning we simply say that the sentence is *strengthened*.)

		Conclusion	
		−strong	+strong
Premise(s)	−strong	Logician	Invalidator
	+strong	Validator	Strengthener

Table 1: Possible reasoner types by the loci of SI computation

Table 1 shows that we can hypothetically distinguish between four types of reasoners. These are reasoners that strengthen (i) neither premises nor conclusions (we call reasoners of this type *Logicians*), (ii) premises and conclusions (*Strengtheners*), (iii) premises but not conclusions (*Validators*), and (iv) conclusions but not premises (*Invalidators*). Reasoners of the first type are called *Logicians* because they only consider the logical relationships between the sentences of an argument. Strengtheners are so called because they strengthen all sentences of an argument. The name *Validator* alludes to the fact that reasoners of this type only strengthen premises, which can only lead to validation of the conclusion.[13] Similarly, the name *Invalidator* relates to the fact that reasoners of this type strengthen the conclusion and only the conclusion, which can only lead to its invalidation.[14]

[11] In PL, A-to-I inferences are not valid. Therefore, some of the rejections of the A-to-I inference might come from subjects that employ PL instead of AL. Importantly, even if there are PL reasoners, the fact that 73% of all subjects *accepted* the A-to-I inference shows that the large majority of subjects did not reject the A-to-I inference on logical grounds (i.e. they are not PL reasoners). Hence, if these subjects accepted the A-to-I inference on logical grounds (i.e. if they are AL reasoners) they did not compute the SI of the I-conclusion.

[12] There are other, perhaps more interesting questions that can be asked at this point. N&G raise the question of why reasoners interpret I-sentences in I-to-O inferences differently than in A-to-I inferences: "The paradox is that the *same* subjects who believe *all* implies *some* also believe that *some* implies the existence of negative instances!" (p. 539 in *op. cit.*) See §6, where we put this issue on the research agenda.

[13] That is, strengthened premises can entail a conclusion which is not entailed by the premises without their SIs.

[14] That is, the conclusion without its SI can be entailed by the premise but the strengthened conclusion may not be entailed.

We put forth the hypothesis that the observed variation in how frequently SIs are computed for premises vs conclusions is due to *individual variation*: there are different reasoning behaviors, i.e., different groups of reasoners.[15] More specifically, we hypothesize that we encounter three groups of reasoners in logical reasoning studies:

(4) The overall population consists of Logicians, Validators, and Strengtheners.

Our hypothesis predicts that I-to-O inferences are accepted by two groups of reasoners, namely by Validators and Strengtheners. Furthermore, it predicts that A-to-I inferences are rejected by just one group, namely by Strengtheners. Thus, it is supported by the observed variation.[16]

To test the hypothesis in (4), we conducted an experiment in which subjects were asked to form a judgment about the logical validity of so-called *syllogisms*, i.e., arguments like those in (5) and (6), where the former is logically valid and the latter logically invalid.

(5)
(A) All Ms are Ks
(I) Some Ps are Ms

(I) Some Ps are Ks

(6)
(A) All Ms are Ks
(E) No Ms are Ps

(O) Some Ps are not Ks

The (in)validity of a syllogism can be affected by SI computation, and their greater complexity allows us to have more variety amongst our experimental items. More importantly, syllogisms can induce more response patterns than arguments with just one premise. Hence, they may yield evidence for all three groups hypothesized in (4) (see §4). The goal of our paper is threefold: (i) to present analytical tools that help quantify the impact of SIs on syllogistic reasoning performance, (ii) to show how these tools can be used to experimentally establish the existence of specific groups of reasoners, and (iii) to discuss to what extent an experiment that we conducted succeeded in doing so.

The paper is structured as follows. We begin with a brief review of (Aristotelian) syllogisms (§2). We then identify six syllogism classes that differ from each other in how SI computation affects (or doesn't affect) their (in)validity (§3). We proceed by spelling out the predictions of the hypothesis that there are three different groups of reasoners, viz. Logicians, Strengtheners, and Validators. That is, we show what response profiles we predict to observe given our

[15] Note that this is not the only possible answer. Other researchers correlate error rates in logical reasoning tasks with processing complexity (e.g. Geurts 2003). Importantly, the processing complexity of a reasoning task is assumed to be the same for all reasoners. For instance, Geurts (2003) proposes a complexity measure assuming an "abstract reasoner."

[16] By being existentials, O-sentences also come with a SI, viz. the SI in (i).

(i) some Ms are not Ks $\overset{\text{SI}}{\rightsquigarrow}$ not all Ms are not Ks

As expected, this SI also affects logical reasoning performance: N&G report that 83% of all subjects accepted logically invalid O-to-I inferences. Again, this can be put down to the fact that the I-conclusion is entailed by the conjunction of the O-premise and its SI. Moreover, it can again be hypothesized that SI computation can lead to rejection of a logically valid inference: in N&G's experiment, E-to-O inferences (i.e., inferences from *no Ms are Ks* to *some Ms are not Ks*), which are valid in AL, were rejected by 31% of all subjects, presumably because of the SI in (i). Note that more reasoners accepted O-to-I inferences than reasoners rejected E-to-O inferences, as predicted by the hypothesis in (4): the former are accepted by Strengtheners and Validators, while the latter are only rejected by Strengtheners.

syllogism classes and these reasoner groups (§4). We then describe the experiment that we conducted and discuss its results (§5). We end with a conclusion and an outlook (§6).

2. Syllogisms

As exemplified in (5) and (6) above, syllogisms are arguments that are made up of three sentences, i.e., two premises and a conclusion. The linguistic form of the sentences of a syllogism as well as their arrangement is subject to restrictions.[17] Every sentence must have one of the following four form types, traditionally called A, I, E, O: (A) all α are β; (I) some α are β; (E) no α are β; (O) some α are not β (where α and β are predicate expressions, henceforth *terms*). The distribution of terms in a syllogism is restricted by two constraints: (i) there is one and only one term – the so-called *middle term* – that occurs in both premises (in (5) and (6), the term *M*); (ii) the unique term of the 2nd/1st premise is the (linearly) 1st/2nd term of the conclusion. Constraint (i) allows four distributions of terms, traditionally called figures: 1. the middle term is the 1st/2nd term of the 1st/2nd premise; 2. the middle term is the 2nd term of both premises; 3. the middle term is the 1st term of both premises; 4. the middle term is the 2nd/1st term of the 1st/2nd premise; for all four cases, constraint (ii) uniquely determines the term distribution in the conclusion. The four figures are graphically represented in (7). The colored boxes represent the terms, which means that blue boxes represent the middle term.

(7)

Figure 1 Figure 2 Figure 3 Figure 4

Consequently, there are 256 syllogisms: 4 sentence types (A, I, E, O) to the exponent of 3 (2 premises + 1 conclusion) × 4 figures. Syllogisms are identified by giving, in this order, the form type of the 1st premise, the form type of the 2nd premise, the figure, and the form type of the conclusion. Thus, (5) is an instance of AI1I and (6) an instance of AE3O.

Of the 256 syllogisms, 24 are valid in AL, and 15 of those are also valid in PL. Valid syllogisms have at least one universal (A or E) premise. The nine syllogisms that are valid in AL but not PL have two universal premises and an existential (I or O) conclusion.[18] There are five valid syllogisms with a universal conclusion, which can only be validated by universal premises.[19] Finally, there are ten valid syllogisms with an existential premise and an existential conclusion.[20] As we will show in §3, the distribution of existential sentences in a syllogism determines its membership in the syllogism classes that we use to test the hypothesis in (4).

[17] We adopt the traditional restrictions from Aristotelian-scholastic logic to make our experimental results more easily comparable with the results of previous studies (e.g. Rips 1994). We agree with Geurts (2003) that these restrictions are mostly arbitrary and hence not particularly interesting from a logical or linguistic point of view.

[18] These are AA1I, AA3I, AA4I, AE2O, AE4O, EA1O, EA2O, EA3O, and EA4O.

[19] These are AA1A, AE2E, AE4E, EA1E, and EA2E.

[20] These are AI1I, AI3I, IA3I, IA4I, AO2O, OA3O, EI1O, EI2O, EI3O, and EI4O.

3. Syllogism classes

We now detail how the ways that SIs can affect the (in)validity of syllogisms define syllogism classes, which form the conditions of our syllogism experiment.[21] We can identify six syllogism classes by how their members are affected by SI computation. The classes are designated by $+v$ if their members are valid in AL, and by $-v$ otherwise. This designation is followed by $\overset{\text{SI}}{\rightsquigarrow} +v$, $\overset{\text{SI}}{\rightsquigarrow} -v$, or $\overset{\text{SI}}{\rightsquigarrow} \pm v$, depending on the effect of SI computation (see below for details). An important outcome of this classification is shown in Table 2. The table foreshadows which reasoner groups of hypothesis (4) are predicted to accept the syllogisms of which of the six classes as valid.

		Effect of SI computation		
		$\overset{\text{SI}}{\rightsquigarrow} -v$	$\overset{\text{SI}}{\rightsquigarrow} +v$	$\overset{\text{SI}}{\rightsquigarrow} \pm v$
Validity status in AL	$-v$	$\emptyset$	Validators Strengtheners	Validators
	$+v$	Validators Logicians	Validators Logicians Strengtheners	Validators Logicians

Table 2: Which groups are predicted to accept which classes

That is, the designations of the syllogism classes also inform about the mappings from validity in AL to the predicted validity judgments of the reasoner groups of hypothesis (4) (see §4).

There are two *invariant* classes, i.e., classes whose members are unaffected by SI computation:

- $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$ Invariantly invalid syllogisms

  The syllogisms in this class are not validated by applying SI computation to their premises. There are three possible reasons for this: (i) SI computation is vacuous (syllogisms without existential premises, e.g. EE3I), (ii) SI computation isn't vacuous but the conclusion can only be validated by two universal premises (syllogisms with a universal conclusion, e.g. IE4E), or (iii) SI computation isn't vacuous but the premises are too weak for the SIs to be able to add enough strength to validate the conclusion (syllogisms with two existential premises, e.g. II4I, OO4I).

[21] There is another kind of non-logical inference that is known to drastically impede logical reasoning performance, viz. *illicit conversion* (IC). By IC, the two terms of an A- or O-sentence are interchanged, see (i).

(i) a. all Ms are Ks $\overset{\text{IC}}{\rightsquigarrow}$ all Ks are Ms

b. some Ms are not Ks $\overset{\text{IC}}{\rightsquigarrow}$ some Ks are not Ms

Note that neither the conversion of the terms of an A-sentence, (ia), nor the conversion of the terms of an O-sentence, (ib), is logically valid (hence the qualification *illicit*). We controlled for this influence on reasoning performance by excluding all syllogisms whose (in)validity is affected by IC. For example, we excluded the logically invalid syllogism AE3O in (6) because it is validated by the IC inference of the A-premise. To give an example of the opposite case, we excluded EI1O because the IC inference of the O-conclusions invalidates this logically valid syllogism.

- $[+v \overset{\text{SI}}{\rightsquigarrow} +v]$ Invariantly valid syllogisms

  The syllogisms in this class are not invalidated by applying SI computation to the conclusion. There is only one possible reason for this: SI computation is vacuous because the conclusion is a universal sentence (AA1A, AE2E, AE4E, EA1E, EA2E).

Furthermore, there are four *variant* classes, i.e., classes whose members are affected by SI computation:

- $[-v \overset{\text{SI}}{\rightsquigarrow} +v]$ Invalid syllogisms that are validated by SI computation

  Since universal conclusions can be only be validated by universal premises, class $[-v \overset{\text{SI}}{\rightsquigarrow} +v]$ can only contain syllogisms with an existential (I or O) conclusion. However, the SI of the existential conclusion must also be validated by the (strengthened) premises, or else SI computation does not necessarily lead to validation. This means that the members of $[-v \overset{\text{SI}}{\rightsquigarrow} +v]$ must be counterparts of a pair of valid syllogisms that differ only in that one contains I-sentences in places where the other contains O-sentences. There is one (and only one) such pair: IA3I and OA3O. This means that $[-v \overset{\text{SI}}{\rightsquigarrow} +v]$ has the following two members (and only these two members): IA3O and OA3I.

- $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ Valid syllogisms that are invalidated by SI computation

  This class contains all valid syllogisms with an existential conclusion (e.g. EI1O), except for the two members of class $[+v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ (see below).

- $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ Invalid syllogisms that are validated by selective SI computation

  Here, "validated by selective SI computation" means that the members of $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ are validated by the strengthened premises but only if the SI of the conclusion is not computed. The class contains invalid syllogisms with an existential premise and an existential conclusion (AO1I, AO3I, OA4I, EO1O, EO2O, EO3O, EO4O). To be a member of $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$, an invalid syllogism must have a valid counterpart in which the existential premise is replaced by its subcontrary (e.g. AI1I for AO1I).

- $[+v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ Valid syllogisms that are invalidated by selective SI computation

  This class contains valid syllogisms with an existential conclusion and an existential premise such that (i) the strengthened conclusion is not entailed by the premises and (ii) the strengthened conclusion is entailed by the SI of the existential premise in conjunction with the other premise. Class $[+v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ has two members, namely IA3I and OA3O.[22]

[22]It can be easily seen that the SI of the I-conclusion of IA3I is entailed by the SI of the I-premise in conjunction with the A-premise: the SIs of the I-conclusion and I-premise are the corresponding O-sentences, and OA3O is a valid syllogism (and the other way around for OA3O).

It can be easily verified that these six classes exhaust the set of syllogisms. However, in section 5.3 we will present refinements of this classification.

4. Predictions

In §1, we formulated the hypothesis that there are three groups of reasoners, which we named Logicians, Validators, and Strengtheners. Table 3 recapitulates how we characterize these groups. In addition to this, the table shows how the members of each group interpret existential premises and conclusions, and what effect this can have for the validity of an argument. Thereby, "weak" stands for the literal 'some or all' meaning of the existential quantifier *some* and "strong" for its 'some and not all' meaning, which is derived by conjunctively adding its SI to the literal meaning. The colors encode the relation between the locus of SI computation and the potential effect of the SI.

	Logicians	Validators	Strengtheners	
	don't compute SIs	compute SIs for premises but not for conclusions	compute SIs for premises and conclusions	
Existential premise	weak	strong	strong	can validate an invalid argument
Existential conclusion	weak	weak	strong	can invalidate a valid argument

Table 3: The three groups and their interpretation of existential sentences of an argument

With the assumption of these three groups of reasoners, we predict to observe three different response patterns in syllogistic reasoning experiments, which are given in Table 4. For obvious reasons, we predict that the validity judgments of Logicians directly reflect the logical validity of a syllogism. That is, they are predicted to accept (✓) the syllogisms of all classes that are designated as valid ($[+v\ldots]$) and to reject (✗) the syllogisms of all classes that are designated as invalid ($[-v\ldots]$). Validators are predicted to accept the syllogisms of all classes that are designated as $[+v\ldots]$, $[\ldots \overset{\text{SI}}{\rightsquigarrow} +v]$, or $[\ldots \overset{\text{SI}}{\rightsquigarrow} \pm v]$ (valid or valid if SI computation applies (only) to the premises), and to reject all others. Finally, Strengtheners are predicted to reject the syllogisms of all classes that are designated as $[\ldots \overset{\text{SI}}{\rightsquigarrow} -v]$ or $[\ldots \overset{\text{SI}}{\rightsquigarrow} \pm v]$ (invariantly invalid or invalid if SI computation applies to the conclusion), and to accept all others.

5. Experiment and results

5.1. The experiment

To test the predictions of our approach, we conducted an experiment with 120 participants over Amazon Mechanical Turk. Since class $[+v \overset{\text{SI}}{\rightsquigarrow} \pm -v]$ evokes the same responses as class $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ for all three groups (see Table 4), we chose not to use tokens of class $[+v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ in the experiment. Each participant was asked to give 100 binary acceptability judgments for

Syllogism class	Logicians	Validators	Strengtheners	
$[-v \overset{\text{SI}}{\rightsquigarrow} -v]$	✗	✗	✗	invariant
$[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$	✗	✓	✗	affected by SI computation
$[-v \overset{\text{SI}}{\rightsquigarrow} +v]$	✗	✓	✓	
$[+v \overset{\text{SI}}{\rightsquigarrow} -v]$	✓	✓	✗	
$[+v \overset{\text{SI}}{\rightsquigarrow} \pm v]$	✓	✓	✗	
$[+v \overset{\text{SI}}{\rightsquigarrow} +v]$	✓	✓	✓	invariant

Table 4: The predicted reasoning patterns for the members of each group

20 tokens of each of the five selected syllogism classes. Participants were told that they will be presented arguments with two premises and a conclusion, and were instructed "to say whether the premises being true means that the conclusion must be true as well." That is, we used a necessity statement and the intuitive notion of the truth of a sentence to evoke judgments about logical validity. The syllogism in (8) exemplifies the tokens that we used in our experiment.

(8)
No Italians are miners
All bikers are Italians

No bikers are miners

For the three terms of the syllogism tokens, we used different nationalities, professions, and hobbies (above *Italian*, *miner*, and *biker*, respectively), without repetitions. Which of the three terms functioned as the middle term was always randomly determined. The experimental task was preceded by a practice session consisting of two arguments that were different in form from (Aristotelian) syllogisms. Participants received feedback to their responses in the practice session.

5.2. Results: acceptance rates

Table 5 shows the mean acceptance rates of the syllogisms of each class. The ordering of the table rows reflects how many groups are predicted to accept the syllogisms of the corresponding class.[23] The table furthermore shows which differences between the mean rates we predict with the hypothesis that there are Logicians (L), Validators (V), and Strengtheners (S): brackets of the right side of Table 5 connect certain pairs of rows; for each bracket (and transitively each connected sequence of brackets), we predict a higher mean acceptance rate for the class at the lower tip of the bracket than for the class at the upper tip.

[23] Since we don't make any predictions about the relative size of the groups, the order of the rows of class $[-v \rightsquigarrow +v]$ and $[+v \rightsquigarrow -v]$ with respect to each other is arbitrary.

Class	L	V	S	% acc.
$[-v \rightsquigarrow -v]$	✗	✗	✗	19
$[-v \rightsquigarrow \pm v]$	✗	✓	✗	56.4
$[-v \rightsquigarrow +v]$	✗	✓	✓	64.6
$[+v \rightsquigarrow -v]$	✓	✓	✗	60.7
$[+v \rightsquigarrow +v]$	✓	✓	✓	76.3

Table 5: Mean acceptance rates and predicted differences

Since the judgments with respect to class $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$ and $[+v \overset{\text{SI}}{\rightsquigarrow} +v]$ are not impeded by SI computation,[24] the error rates of 19% false positives and 23.7% false negatives are the most immediate reflexes of true performance errors. The error rates show that there is no general positive response bias. Five of the six predictions marked in Table 5 are borne out. However, ANOVA and post-hoc tests show that the difference between the mean acceptance rate of class $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ and $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ does not reach significance. That is, our prediction that syllogisms of class $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ are accepted more often than syllogisms of class $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ because the former are accepted by Logicians and Validators, while the latter are only accepted by Validators, is *not* borne out.

5.3. Discussion

As was just pointed out, the difference between the mean acceptance rate of class $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ and $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ does not reach significance. On closer inspection, the reason for this is that there is too much variation in acceptance rates across the syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$. For instance, the tokens of the type in (9) are accepted $\sim 80\%$ of all times, while the tokens of the type in (10) are only accepted $\sim 50\%$ of all times.

(9)
(A) All Ms are Ks
(I) Some Ms are Ps

(I) Some Ps are Ks

(10)
(A) All Ks are Ms
(E) No Ms are Ps

(O) Some Ps are not Ks

As can be easily seen,[25] the latter syllogisms are only valid in AL, while the former are valid in both AL and PL. That is, we observe that in $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ syllogisms that are valid in both AL and PL are accepted more often than syllogisms that are only valid in AL.[26, 27]

[24] Note also that we excluded syllogisms whose (in)validity is affected by IC inferences (see footnote 21).

[25] Recall that all syllogisms with two universal premises and an existential conclusion, such as AE4O and EA3O, are invalid in PL and that all syllogisms that are valid in PL are also valid in AL.

[26] We did not collect the information whether a participant had training in formal logic. However, Rips (1994) notes that the subjects of his and Jeffrey Schank's experiment were "20 University of Chicago students, none of whom had taken a course in logic." Importantly, the data set of Rips (1994) also suggests that a syllogism's validity in PL is a relevant factor for the acceptance rates within class $[+v \rightsquigarrow -v]$: syllogisms in $[+v \rightsquigarrow -v]$ that are valid in both AL and PL were accepted 68% of all times and syllogisms in $[+v \rightsquigarrow -v]$ that are only valid in AL 51% of all times.

[27] One might think that this result is expected since Rips (1994) already notes that "subjects gave 85.8% "follows"

Another distinction that we overlooked in the design of our experiment is whether or not the sentences of a syllogism are inconsistent (before or after strengthening). Taking inconsistency into account leads to the following subclassifications of the classes identified in §3, where the designation $-c$ stands for 'inconsistent:'

- Subclass $[-v \overset{\text{SI}}{\rightsquigarrow} -c]$ of $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$

  Class $[-v \overset{\text{SI}}{\rightsquigarrow} -c]$ contains syllogisms with the following properties: the SI of one of its premises in conjunction with the other premise entails the contradictory of the conclusion. The syllogisms in this class (e.g. AI2A) have a valid syllogism as a counterpart which expresses the problematic entailment.[28]

- Subclass $[-c]$ of $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$

  This class contains syllogisms that are formed from sets of inconsistent sentences, i.e., counterparts of valid syllogisms in which the valid conclusion is replaced by its contradictory (e.g. AA1O, which is the inconsistent counterpart of the valid syllogism AA1A).

- Subclass $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$ of $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$

  Class $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$ contains all valid syllogisms with an existential (I or O) conclusion that have a valid counterpart in which the superaltern (A or E) is the conclusion (e.g. AA1I, which has AA1A as a counterpart; the SI of the I-conclusion of AA1I is the contradictory of the A-conclusion of AA1A).

The relevance of this subclassification for syllogistic reasoning studies can be seen from the fact that the rate of false positives is lower for class $[-v \overset{\text{SI}}{\rightsquigarrow} -c]$ and class $[-c]$ than for class $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$, where $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$ is now taken to exclude the syllogisms in the former two classes (i.e. the designation $-v$ now stands for 'invalid but consistent'). This is shown in Table 6.[29]

Class	% acc. in Rips (1994)
$[-v \overset{\text{SI}}{\rightsquigarrow} -v]$	10.3
$[-v \overset{\text{SI}}{\rightsquigarrow} -c]$	1.5
$[-c]$	1

Table 6: The effect of inconsistency on the rate of false positives

responses to [syllogisms that are valid in both AL and PL], but only 63.3% "follows" responses to the nine [syllogisms that are only valid in AL]. In Dickstein's study [Dickstein 1978], the corresponding percentages are 89.4 and 70.8." Importantly, however, the class of syllogisms that are valid in both AL and PL has class $[+v \rightsquigarrow +v]$ as a subclass. The syllogisms in $[+v \rightsquigarrow +v]$ are expected to be accepted more often than any other syllogism (and hence to lift the mean acceptance rate of its superclass) because they are logically valid and not invalidated by SI computation. From our point of view, it is highly unexpected to find an effect of PL validity *within* class $[+v \rightsquigarrow -v]$.

[28]In the case of AI2A, the valid counterpart is AO2O: its O-premise is the SI of the I-premise of AI2A and its O-conclusion is the contradictory of the A-conclusion of AI2A.

[29]Our item set does not contain tokens of class $[-v \overset{\text{SI}}{\rightsquigarrow} -c]$ or $[-c]$. Therefore, we use the data of Rips (1994) in Table 6.

There are two possible explanations of the effect of inconsistency on the rate of false positives: (i) inconsistency leads to better recognition of invalidity; (ii) there are reasoners that do not form a judgment about logical consequence but about logical consistency (i.e. they check whether the conclusion is logically consistent with the premises).

We do not observe the same effect of inconsistency on false negatives. That is, as shown in Table 7 the rate of false negatives is not lower in class $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$ than in class $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$.[30]

Class	% acc. in our data	% acc. in Rips (1994)
$[+v \overset{\text{SI}}{\rightsquigarrow} -v]$	53.4	51.3
$[+v \overset{\text{SI}}{\rightsquigarrow} -c]$	51.9	57

Table 7: The effect of inconsistency on the rate of false negatives

In the context of our hypothesis that there are Strengtheners, this result neither supports hypothesis (i) nor hypothesis (ii). For both syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ and syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$, Strengtheners compute the SI of the conclusion. By hypothesis (i), they recognize the resulting invalidity of the logical consequence relation better for syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$ than for syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$. That is, by hypothesis (i) they are predicted to reject syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$ more often than syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$. By hypothesis (ii), some of the Strengtheners may form a consistency judgment instead of a judgment about logical consequence. Therefore, by hypothesis (ii) they are predicted to accept syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ more often than syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$. Neither prediction is supported by the observed data. A possible explanation for the data in Table 7 is that the SI of an existential conclusion is (sometimes) not computed if the premises settle the stronger universal alternative (i.e. if they entail the universal alternative or entail its negation).[31] In the case of the syllogisms in $[+v \overset{\text{SI}}{\rightsquigarrow} -c]$, the universal alternative is settled by the premises since they entail the contradictory of the SI of the conclusion, which is the negation of the universal alternative.

5.4. Results: identifying groups of reasoners

In this section, we illustrate how to determine whether the observed mean acceptance rates reflect homogeneous behavior within different groups and not heterogeneous behaviour of a single group (i.e. all subjects). Recall that every participant of our experiment gave a judgment about 20 tokens of each of the five selected syllogism classes. This means that for every participant we have a rich response profile by means of which we can detect consistent behavior

[30] All of the syllogisms in class $[+v \rightsquigarrow -c]$ are invalid in PL, since they have two universal premises and an existential conclusion. Therefore, the numbers for class $[+v \rightsquigarrow -v]$ in Table 7 reflect only the acceptance rates of PL-invalid syllogisms. Our data set contains tokens of only one such syllogism. The apparent difference between the acceptance rate of the syllogism in $[+v \rightsquigarrow -v]$ and the mean acceptance rate of class $[+v \rightsquigarrow -c]$ is not significant.

[31] According to Fox (2007), SI computation is motivated by the goal to reduce speaker ignorance inferences. If the premises of a syllogism settle the universal alternative of the conclusion, no speaker ignorance inference arises for this alternative and hence there is no motivation to derive the SI of the conclusion.

of individuals and similarities in behavior between individuals. To identify subpopulations in our data set, we used a density-based clustering algorithm, DBSCAN (Ester et al., 1996). With DBSCAN, a density cluster is defined by specifying what counts as a populated neighborhood of a data point (viz. by specifying how many data points must be minimally within a specified radius around that point). Density clusters consist of core points and border points. A data point is a core point if it has a populated neighborhood; a data point is a border point if it is in the neighborhood of a core point but not itself a core point; all other data points are outliers.

The behavior towards the two invariant classes, $[+v \overset{\text{SI}}{\rightsquigarrow} +v]$ and $[-v \overset{\text{SI}}{\rightsquigarrow} -v]$, gives a measure of a subject's logical abilities. This measure can be used to gauge the subject's behavior towards the three variant classes by how much it deviates from the subject's logical abilities.

Since there are three variant classes, the subjects' reasoning behavior towards the variant classes can be mapped into a three-dimensional coordinate space, which is shown in Figure 1. Perfect Logicians are mapped onto the front lower right corner. The distance from this corner along the three dimensions represents how much a subject deviates from a perfect Logician. Perfect Validators deviate maximally from perfect Logicians along two dimensions, the x-dimension, on which deviance towards class $[-v \overset{\text{SI}}{\rightsquigarrow} +v]$ is represented, and the z-dimension, on which deviance towards class $[-v \overset{\text{SI}}{\rightsquigarrow} \pm v]$ is represented. Perfect Strengtheners also deviate maximally from perfect Logicians along two dimensions, the x-dimension and the y-dimension, on which deviance towards class $[+v \overset{\text{SI}}{\rightsquigarrow} -v]$ is represented. Other corners can also be characterized in terms of the reasoning behavior that a subject must have to be mapped onto that corner. The corner that is opposite of the Validators' corner along all three dimensions is the Invalidators corner. Subjects in this corner compute SIs for conclusions but not for premises. The corner that is opposite of the Logicians' corner is Mephistopheles' corner. Like Mephistopheles, subjects in this corner always negate.

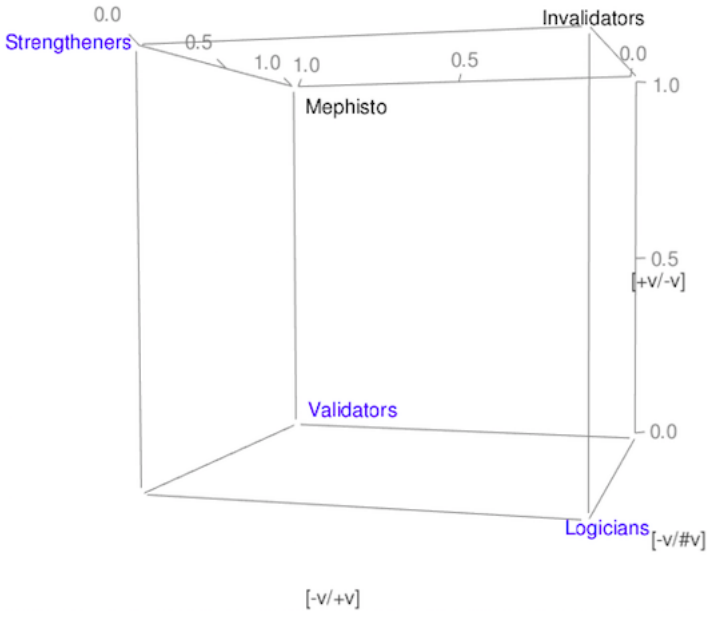

Figure 1: The coordinate space into which the subjects' reasoning behavior is mapped

In Figure 2 and Figure 3, we show the inhabited coordinate space, i.e., the space onto which all subjects with an error rate of $\leq 12.5\%$ relative to the invariant classes are mapped ($\approx$ half of all subjects).

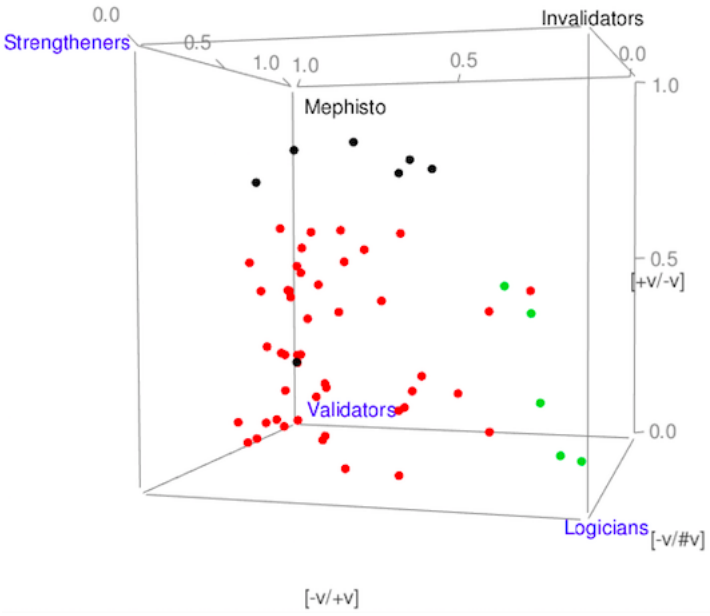

Figure 2: The inhabited coordinate space

The figures show that there are two density clusters (determined by DBSCAN), which are marked in red and green (outliers are black). The Logicians' corner is a border point of the green cluster (i.e., it is in the neighborhood of a core point of the green cluster but not itself a core point). This means that the subjects that belong to the green cluster can count as Logicians. Similarly, the Validators' corner is a core point of the red cluster. Hence, the subjects that belong to the red cluster can count as Validators.[32] The two perspectives provided by Figure 2 and Figure 3 show that the Strengtheners' corner is not populated and neither is any other corner. This means that there is no evidence for populations other than Logicians and Validators. Note that Figure 3 shows that almost all subjects are above the zero point of the z-axis and Figure 3 shows that almost all subjects are left of the diagonal of the base square of the cube. This means that almost all subjects strengthen conclusions sometimes. However, we don't observe systematic strengthening of conclusions, i.e., there are no Strengtheners.

In a certain sense, our data do not contain a lot of noise: as the result of DBSCAN shows our data set contains only few outliers. However, we still need to be concerned about the quality of the data since the clusters that we can identify and associate with specific reasoning behaviors are very spacious. That is, the large majority of points are very distant from the corners that represent the reasoning behaviors that we hypothesized to exist. This means that only a small

[32]Since the number of density clusters and their size depend, by design, on the parameter settings that determine what counts as a populated neighborhood, different parameters would have produced different results. The point of our demonstration is to show that there are parameters that determine two clusters that we can identify with two reasoner groups of hypothesis (4). Importantly, there is no parameter setting that would give us the group of Strengtheners of the group of Invalidators.

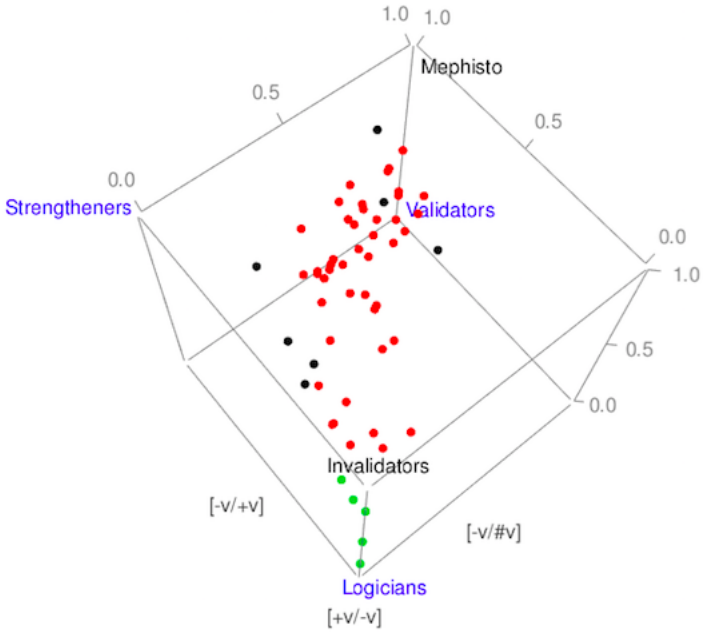

Figure 3: Another perspective on the inhabited coordinate space

proportion of subjects showed the hypothesized behaviors consistently. We think that this is a consequence of the experimental design (primarily the length of the experiment).

6. Conclusion and outlook

We have presented a classification of syllogisms that allows to quantify how frequently the premises and/or the conclusion of a syllogism is strengthened by SI computation. Furthermore, we have put forward the hypothesis that there are three groups of reasoners, viz. Logicians, Validators, and Strengtheners, whose reasoning behavior is characterized in terms of the loci of SI computation in logical arguments. In this way, we could argue that the variation in error rates observed across syllogisms is an effect of individual variation: members of different reasoner groups form the same judgment for the syllogisms of some classes and different judgments for the syllogisms of other classes. The experimental results that we presented support this hypothesis to a certain extent. For instance, the assumption that there is a group of Strengtheners makes correct predictions for the mean error rates of certain classes. Problematically, though, there is no further evidence for the hypothesized group of Strengtheners. That is, our data set contains no cluster of response patterns that can be identified with the response pattern of an idealized Strengthener. Importantly, however, we *did* find this kind of evidence for the two other groups, viz. for the groups of Logicians and Validators.

In future research, we want to address two issues: (i) There is evidence that suggests that some reasoners employ PL in syllogistic reasoning tasks. This behavior raises the question under what circumstances these reasoners associate natural language quantifiers with non-Aristotelian meanings. Answering this question will inform us about the nature of the requirement that ensures the Aristotelian property of (strong) natural language quantifiers, i.e., the requirement that the restriction of a quantifier be non-empty. (ii) The reasoning behavior of Validators shows

the preferred locus for SI computation in logical arguments, viz. the premises of an argument. We want to answer the question whether such a preference can also be found in other supra-sentential contexts, and if so, how these contexts can be characterized. This will inform us about the reason why Validators employ SI computation selectively and hence inform about the motivation for SI computation being employed in natural language discourse.

References

Begg, I. and G. Harris (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior 21*, 595–620.

Chierchia, G., D. Fox, and B. Spector (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, and K. von Heusinger (Eds.), *Semantics: An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter.

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory and Cognition 6*, 76–83.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231. AAAI Press.

Fox, D. (2007). Free choice disjunction and the theory of scalar implicatures. In U. Sauerland and P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics*, pp. 71–120. Houndmills: Palgrave-Macmillan.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition 86*, 223–251.

Grice, P. (1975). Logic and conversation. In D. Davidson and G. Harman (Eds.), *The Logic of Grammar*, pp. 64–75. Encino, CA: Dickenson.

Newstead, S. E. and R. A. Griggs (1983). Drawing inferences from quantified statements: A study of the square of oppositions. *Journal of Verbal Learning and Verbal Behavior 22*, 535–546.

Rips, L. J. (1994). *The Psychology of Proof*. Cambridge, Mass.: MIT Press.

Sells, S. B. (1936). The atmosphere effect: An experimental study of reasoning. *Archives of Psychology 29*, 3–72.

Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology 16*, 1–83.

Woodworth, R. S. and S. B. Sells (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology 18*, 451–460.