When logical conclusions go against beliefs: an ERP study.

Running title: Brain responses to categorical conclusions

Authors: Pablo Rodríguez-Gómez[1,2], Irene Rincón-Pérez[1,2], Gerardo Santaniello[1], Claudia Poch[1,2], Miguel A. Pozo[1], José A. Hinojosa[1, 2], and Eva M. Moreno[1]*

[1]Human Brain Mapping Unit, Instituto Pluridisciplinar, Universidad Complutense de Madrid, Paseo de Juan XXIII 1, 28040, Madrid, Spain

[2]Facultad de Psicología, Campus de Somosaguas, Universidad Complutense de Madrid, 28223, Madrid, Spain

* Correspondence should be addressed to:

Eva M. Moreno, Human Brain Mapping Unit, Instituto Pluridisciplinar, Universidad Complutense de Madrid, Paseo Juan XXIII 1, 28040 Madrid, Spain

Tel.: +34 91 394 32 61; fax: +34 91 394 32 64

E-mail address: emmoreno@ucm.es

ABSTRACT

Reasoning is a fundamental human ability, vulnerable to error. According to behavioral measures, we are biased to consider valid the conclusion of an argument based on the veracity of the conclusion itself rather than on the formal logic of the argument. Nowadays, brain imaging techniques can be used to explore people´s responses as they reason with linguistic materials. Using the Event-Related Potential technique in a categorical syllogism reading task, an N400 enhancement was found for the processing of invalid conclusions preceded by true premises (e.g. *All men are mortal*). By contrast, when initial premises consisted of socially prejudiced statements previously rated as false (e.g. *All blond girls are dumb*), valid rather than invalid conclusions enhanced the N400 response. Considering what the modulation of N400 indexes (i.e., word anticipation processes), our data suggests that people cannot follow the logic of an argument to anticipate upcoming words if they clash with veracity.

Keywords: Event-related potentials; Language; Categorical reasoning; N400

1. Introduction

Reasoning is a fundamental human ability and language is the instrument of thought. Initial statements (premises) guide our reasoning leading us to valid or invalid conclusions. Whenever it follows a correct reasoning form, an argument is said to be *valid*. Otherwise, it is an invalid argument. Additionally, to reach to a valid and also *sound* conclusion we need yet another basic ingredient: the conclusion must be drawn from *true* rather than false premises.

1.1. ERP studies on conditional reasoning

Nowadays, brain imaging techniques can be used to explore people´s responses as they reason with linguistic materials. Remarkably, different types of arguments (e.g. categorical, propositional, transitive) have been shown to engage distinct brain networks (Prado, Chadha, & Booth, 2011). Among brain imaging techniques, Event-Related Potentials (ERPs) have an excellent time resolution of the order of milliseconds (ms). Experimental manipulations in the attentional, linguistic, or memory domain lead to amplitude increases/decreases of specific ERP components (e.g. P2, P3, N400), each of which is linked to distinctive cognitive processes. Thus, for example the N400 ERP component, peaking at around 400 milliseconds, has been linked to the ease of integration/prediction of upcoming words in discourse (Kutas & Federmeier, 2011). The smaller the amplitude of the component, the more expected the word ending of the sentence was.

To date only a few ERP studies have been conducted on deductive reasoning, mostly examining conditional reasoning. This form of reasoning involves arguments in the form of: If p then q, p, then q (Blanchette & El-Deredy, 2014; Bonnefond & Henst, 2013; Bonnefond, Kaliuzhna, Van der Henst, & De Neys, 2014; Bonnefond & Van der Henst, 2009; Qiu et al., 2007). In arguments such as: (1) If a figure is a square then it is red. (2) The figure is a square. (3) Therefore it is red/a square (inference vs. repetition conditions), increased negative amplitudes were elicited in the 500-700 time window and later between 1700 and 2000 ms by

both logically valid and invalid inferences relative to the repetition of the minor premise (Qiu et al., 2007). The earlier effect was related to the activation and the application of the inference rules, whereas the later was suggested to reflect cognitive effort to verify whether the deductive conclusions were correct. Thus, the making of an inference regardless of its validity has been linked to late ERP components. It is important, however, to point out that ERPs in this study were measured time-locked to the onset of the second minor premise and not to the processing of the conclusion itself.

Bonnefond et al. (2014) explored reasoning with more realistic semantic conditionals, such as: (1) If the butter warms then it melts. (2) The butter warms. (3) The butter melts. Their study considered that some conditions might prevent the consequent from occurring despite the presence of the antecedent, such as in: `If John studies hard, he would pass the test´. The reader can take into consideration some "disablers" for this argument: John might have a low IQ; the test could be very hard; John might not make it to the exam. Whether few or many disablers can be produced in a prior written task becomes critical. Responses to conditionals reveals a different ERP pattern at the point of the conclusion. In particular, an enhanced frontal N2 and a reduced parietal P3b effect were elicited in response to conditionals with many disablers relative to conditionals with fewer disablers, indicating that the many disablers condition made conditional conclusions harder to process. In particular, these effects were linked to the violation and satisfaction of expectations, respectively.

Finally, relative to the repetition of the premises, larger P3b and N400 effects were obtained for inference making conditions in the study by Blanchette & El-Deredy (2014). Conclusions were presented either with 2-5 words simultaneously in the screen (experiment 1) or in response to single words as follows: If dead, then morgue. Morgue. Dead. (experiment 2). According to their results, the authors concluded that inferences were drawn spontaneously before the conclusion was presented. However, in line with previous ERP studies that show an

N400 reduction to words that could have been expected in previous context, a reduction rather than an increase should have been obtained for words potentially inferred from previous reading. The fact that the contrast was made in relation to the repetition of the minor premise may be crucial since word repetition has been associated with an N400 reduction (Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991).

Summarizing, a limited set of studies has been conducted so far on reasoning using the ERP technique and it is difficult to integrate the existing work; every study addressed a slightly different question, examined different components, or investigated different types of reasoning, at different stages. Mostly, conditional reasoning is explored and the baseline condition is the repetition of the minor premise. Considering what the N400 ERP component indexes, we set out to explore online reasoning in the realm of categorical conclusions, where word anticipation processes are most likely to occur.

1.2. Categorical syllogisms

Categorical syllogisms are a form of deductive reasoning (Striker, 2009). They consist of three parts: a major premise, a minor premise, and a conclusion. According to philosophers, a *valid* conclusion is achieved if the syllogism *form* is correct. For example, assuming that all men are mortal (a general statement or major premise) and that Peter is a man (a particular statement or minor premise) one may validly conclude that Peter is mortal (conclusion). That conclusion does not come as a surprise and it might be anticipated by readers if the logical thread is followed. The previous example is a, so-called, affirmative "DARII" type of syllogism, since its major premise is a universal affirmative proposition (symbolized as A), while the minor premise and the conclusion are particular affirmative propositions (symbolized as I). DARII syllogisms are one of the 24 logically valid types of syllogisms, others being named: BARBARA, BOCARDO, FERIO, etc. The DARII example above, besides being a logically *valid* form of syllogism, is also a *sound* argument as both the major and minor premises happen to be

true. Syllogisms, however, can be logically valid (the conclusion logically follows from the premises) but their premises may be uncertain or even false, in which case logically valid but *unsound* conclusions are attained (i.e. valid conclusions based on false premises). A critical question is whether our brains strictly follow the logic of arguments to predict upcoming words in discourse or are also influenced by the veracity of the premises during online reasoning tasks.

So far, behavioral research on categorical thinking has been committed to the understanding of the kind of errors people commit when reasoning with syllogisms. For example, the hypothesis of the *illicit conversion* of the premises (Chapman & Chapman, 1959) posits that people erroneously assume that "*All A´s are B´s*" is the same as "*All B´s are A´s*". Another source of error is called the *belief-bias effect*, according to which conclusions that seem likely are considered to be valid, irrespective of their logical validity (J. S. Evans, J. L. Barston, & P. Pollard, 1983). Two opposing theories were put forward: The rationality theory (Revlin, Leirer, Yopp, & Yopp, 1980) posed that people always follow logical rules and the errors they made arise at the wrong encoding of the premises. Thus, researchers claimed that we logically reason but we do so upon materials whose premises have been illicitly converted. In contrast, Evans et al. (1983) claimed that the believability in the conclusion effect still arises even in conditions in which the conversion of the premises is controlled for. According to this view, errors are still committed due to the strength of the veracity or falsity of the conclusions. Yet another theory, the mental model theory of reasoning (Johnson-Laird, 1975), posits than individuals grasp that an inference is no good if there is a counterexample to it (cited in Khemlani & Johnson-Laird, 2012). Based on the later theory, major premises might be crucial, i.e., especially when they are false universal premises (e.g. All Xs are Y). Nonetheless, a recent meta-analysis of the theories of syllogism concludes that none of them provides an adequate account of syllogistic reasoning errors (Khemlani & Johnson-Laird, 2012). A model based on formal rules of inference for human reasoning might be unable to account for recent evidence

on how we process information. Human everyday reasoning, is best viewed as solving probabilistic, rather than logical, inference problems (Oaksford & Chater, 2009)

1.3. Our ERP study on categorical reasoning

When reasoning occurs with categorical syllogisms, the conclusion for a valid and a sound argument could most likely be anticipated before it appears on the screen (e.g. All men are mortal. Juan is a man. Therefore, Juan is *mortal*). Here, the response to the last word of the conclusion is critical. Since the N400 effect is an index of word anticipation processes at a semantic level (Federmeier, 2007), it is expected that the more anticipated the word, the smaller the N400 amplitude associated with it. Other language related ERP components (e.g. ELAN, LAN, P600) are, by contrast, related to expectations at a syntactic level or are indexing sentence reanalysis processes. Critically, we manipulated the value of truth (belief) assigned to the major premise (All men are mortal *versus* All blond girls are dumb) in an attempt to elucidate whether the logic of the argument was still followed even when the major premise was considered false and stereotyped.

In the ERP field, the seminal work by Kutas & Hillyard (1980) found that words that render statements senseless (e.g. He spread the warm bread with *socks*.) elicited a large negative-going voltage at the brain scalp at around 400 ms (i.e. the N400) from the onset of the critical word *`socks´*. Since then, the N400 proves to be sensitive (i.e. larger in amplitude) to words embedded in context that: 1) make untrue statements such as: "The Dutch trains are *white* and very crowded." when they are in fact yellow (Hagoort, Hald, Bastiaansen, & Petersson, 2004a), and 2) make statements clash with our own beliefs, e.g. "I think the increasing emancipation of women is a *negative* development" (J. J. A. Van Berkum, B. Holleman, M. Nieuwland, M. Otten, & J. Murre, 2009). Target words in previous examples are highlighted in italics. In recent years, the N400 is viewed as an index of a facilitatory process

that takes place for words in context that are either anticipated or easier to semantically be integrated in their context (DeLong, Urbach, & Kutas, 2005; Federmeier, 2007; Kutas & Federmeier, 2011).

Our study explored the N400 phenomenon in the realm of categorical syllogisms. Valid and invalid syllogisms were presented following premises previously rated in veracity terms (as true or false major premises).

With regard to syllogisms with true premises, if participants are able to use the premises to anticipate a conclusion, we expect a smaller N400 to logically valid and sound conclusions (i.e. to the word `*mortal´* in "Therefore, Juan is *mortal*" when preceded by "All men are mortal" and "Juan is a man") than to invalid conclusions (e.g. to the word `*man´* in "Therefore, Juan is a *man*" after reading "All men are mortal" and "Juan is mortal"). However, if the hypothesis of the illicit conversion holds true, i.e., "All men are mortal" is erroneously taken as "All mortal are men", participants will have a difficulty to detect that the conclusion drawn in the second example is invalid, in which case they will not elicit an N400. On the contrary, if they are able to detect the misleading logic of the argument, they will elicit an N400 to invalid conclusions. However, as this is the first ERP study manipulating these variables, other ERP components might be altered at this point in the conclusion (e.g. a late positivity or P600).

With regard to syllogisms with false premises, a critical condition is the amplitude of the N400 elicited by valid-yet-*unsound* conclusions. If participants are able to follow the instruction to be solely guided by the logic of the argument, disregarding major premises veracity, they will anticipate the valid yet *unsound* conclusion (such as: Therefore, Raquel is *dumb*, after reading: All blond girls are dumb. Raquel is blond.) Anticipation would thus prevent the elicitation of an N400 response. In contrast, if the lack of veracity of the major premise is rapidly taken into consideration, as prior studies on the processing of false or morally unacceptable statements

show (Hagoort et al., 2004a; J. J. A. Van Berkum et al., 2009), participants will instead elicit an N400 to this perfectly valid yet *unsound* conclusion, thus indexing that despite the logical form of the argument, an anticipation was not carried out whenever the major premise was untrue.

Methods

Participants

A sample of twenty-nine native Spanish speakers (9 males, mean age = 22.5 years, range = 18-48 years) volunteered to participate in the study in exchange for course credits. All participants gave written informed consent. Twenty-seven participants reported being right handed. The average handedness score (Oldfield, 1971) was 60.8. All participants reported normal or corrected-to-normal vision and none had a history of neurological or psychiatric disorders. No participant data was excluded from analysis based on these exclusion criteria.

2.2. Materials

An initial set of experimental stimuli was created. It consisted of 280 major premises, all written in Spanish. Half of these major premises (140) began with the word “Todos” (“Todos” = All), and the other half began with the word “Ningún” (“Ningún” = No). The structure of the major premises was thus mixed in order to avoid automaticity in the responses from participants, which had to randomly alternate between affirmative (DARII) and negative (FERIO) types of syllogism. The major premises were elaborated with the aim that most people would tend to categorize them as true or false. They consisted of universal affirmative or negative propositions (e.g. All men are mortal; No obese is thin; All blond girls are dumb; No obese can be happy) (see Appendix for a full list of major premises used as stimuli).

These major premises lead to two logically *valid* forms of syllogisms: DARII in the case of affirmative propositions and FERIO in the case of negative propositions, as follows:

Logically Valid		Affirmative (DARII)		Negative (FERIO)	
True Premises	Major premise	All men are mortal.	Universal Affirmative (A)	No obese is thin.	Universal Negative (E)
	Minor premise	Juan is a man.	Particular Affirmative (I)	Coral is obese.	Particular Affirmative (I)
	Conclusion	Therefore, Juan is a mortal.	Particular Affirmative (I)	Therefore, Coral is not thin.	Particular Negative (O)
False Premises	Major premise	All blond girls are dumb.	Universal Affirmative (A)	No obese can be happy.	Universal Negative (E)
	Minor premise	Raquel is blond.	Particular Affirmative (I)	Raúl is obese.	Particular Affirmative (I)
	Conclusion	Therefore, Raquel is dumb.	Particular Affirmative (I)	Therefore, Raúl cannot be happy.	Particular Negative (O)

Table 1. Logically valid DARII and FERIO syllogisms with true and false premises.

The examples in the first rows are valid and sound arguments because 1) they follow a correct form of syllogism; and 2) both premises are true. In contrast, the examples in the last row are logically valid (they follow a correct syllogism form) yet *unsound* arguments because their conclusion originates from false premises.

As a second step, in order to create similar but logically *invalid* syllogisms we used the *fallacy of the undistributed middle term* by switching the minor premise and the conclusion as follows:

Logically invalid		Affirmative		Negative	
True Premises	Major premise	All men are mortal.	Universal Affirmative (A)	No obese is thin.	Universal Negative (E)
	Minor premise	Juan is a mortal.	Particular Affirmative (I)	Coral is not thin.	Particular Negative OI)
	Conclusion	Therefore, Juan is a man.	Particular Affirmative (I)	Therefore, Coral is obese.	Particular Affirmative Negative (I)
False Premises	Major premise	All blond girls are dumb.	Universal Affirmative (A)	No obese can be happy.	Universal Negative (E)
	Minor premise	Raquel is dumb.	Particular Affirmative (I)	Raúl cannot be happy.	Particular Negative (OI)
	Conclusion	Therefore, Raquel is blond.	Particular Affirmative (I)	Therefore, Raúl is obese.	Particular Affirmative (OI)

Table2. Logically invalid syllogisms (fallacies) with true and false premises.

Deductive fallacies like the ones presented above, fail in the transition from general statements to specific instances. The fact that Juan is a mortal does not necessarily imply that Juan is a man. Based on the hypothesis of the illicit conversion (Chapman & Chapman, 1959),

these invalid syllogisms rely on the common error of misinterpreting the major premise (that is, to take “All men are mortal” to be the same as “All mortals are men”). If that was the case, the invalid conclusion that “Juan is a man” would be taken as valid.

For major premise veracity rating purposes, all major premises were divided into four lists (70 major premises per list) and were subjected to subjective evaluation. A hundred subjects (25 per list) were asked to evaluate the major premises included in a list in terms of their veracity. Based on the idea that there are few propositions that people can hold as certainly true, or certainly false because of “most of our beliefs come in degrees” (Evans, Thompson, & Over, 2015), we used a five point rating scale. Participants chose between five different options, as follows: The statement is: 1, totally false; 2, partially false; 3, neither true nor false; 4, partially true; 5, totally true. According to the mean value of truth obtained using this procedure, all major premises were then divided into four different groups. The first group consisted of a priori “true” major premises beginning with “All” (range of value of truth = 1.48 – 5), the second group comprised “true” major premises starting with “No” (range of value of truth = 1.92 – 4.96), the third group consisted of “false” major premises beginning with “All” (range of value of truth = 1.24 – 3.72), and the fourth group comprised “false” major premises starting with “No” (range of value of truth = 1.04 – 4.6). Before the ERP experiment was setup, 30 major premises of each group were rejected based on their value of truth, not being considered highly true or highly false. The major premises with the higher value of truth were kept for groups one and two, and the major premises with the lower value of truth were kept for groups three and four, for a total of 40 sentences per group. The final values of truth for each group were as follows: group 1 (mean = 4.34; range = 3.6 – 5), group 2 (mean = 4.51; range = 4.12 – 4.96), group 3 (mean = 1.59; range = 1.24 – 1.92), group 4 (1.54; range = 1.04 – 1.84). This procedure was followed to ensure that the major premises within each group were considered as *true* or as *false* as possible.

Following a logical sequence, every major premise was continued with a minor premise and a conclusion. The major premise was always a general statement and the minor premise was a particular case related to this general statement. For example: All men are mortal (major premise); Juan is a man (minor premise); Therefore, Juan is mortal. Since we aimed to investigate brainwave responses to valid and invalid conclusions, in half of the syllogisms the order of the minor premise and the conclusion was swapped to obtain invalid conclusions. For example: All men are mortal (major premise); Juan is a mortal (minor premise); Therefore, Juan is a man (conclusion). The later would be an invalid conclusion, since the fact that "All mean are mortal" does not imply that "All mortals are men". In the previous example, Juan could be an unhuman mortal (e.g. an animal whose nickname is Juan). Thus, the conclusion that "Juan is a man" is not warranted.

Finally, syllogisms were divided into eight different groups, based on the veracity of the major premise (true or false), the type of syllogism (Starting with "All" or "No", corresponding to DARII and FERIO type of syllogisms, respectively), and the validity of the conclusion (logically valid or invalid). Table 4 exemplifies each of these group of syllogisms.

Table 4. Groups of syllogisms presented in the experiment in Spanish and their English translation (in italics).

Group	Major premise veracity	First word	Major premise	Minor premise	Conclusion	Valid/invalid conclusion	Value of truth (mean)
1	True	Todos	Todos los hombres son mortales.	Juan es un hombre.	Por tanto, Juan es mortal.	Valid	4.34
1	*True*	*All*	*All men are mortal.*	*Juan is a man.*	*Therefore, Juan is a mortal.*	*Valid*	*4.34*
2	True	Todos	Todos los hombres son mortales.	Juan es mortal.	Por tanto, Juan es un hombre.	Invalid	4.34
2	*True*	*All*	*All men are mortal.*	*Juan is mortal.*	*Therefore, Juan is a man.*	*Invalid*	*4.34*
3	True	Ningún	Ningún obeso es delgado.	Coral es obesa.	Por tanto, Coral no es delgada.	Valid	4.51
3	*True*	*No*	*No obese is thin.*	*Coral is obese.*	*Therefore, Coral is not thin.*	*Valid*	*4.51*
4	True	Ningún	Ningún obeso es delgado.	Coral no es delgada.	Por tanto, Coral es obesa.	Invalid	4.51
4	*True*	*No*	*No obese is thin.*	*Coral is not thin.*	*Therefore, Coral is obese.*	*Invalid*	*4.51*
5	False	Todas	Todas las rubias son tontas.	Raquel es rubia.	Por tanto, Raquel es tonta.	Valid	1.59

5	*False*	*All*	*All blonde girls are dumb.*	*Raquel is blonde.*	*Therefore, Raquel is dumb.*	*Valid*	*1.59*
6	False	Todas	Todas las rubias son tontas.	Raquel es tonta.	Por tanto, Raquel es rubia.	Invalid	1.59
6	*False*	*All*	*All blonde girls are dumb.*	*Raquel is dumb.*	*Therefore, Raquel is blonde.*	*Invalid*	*1.59*
7	False	Ningún	Ningún obeso puede ser feliz.	Raúl es obeso.	Por tanto, Raúl no puede ser feliz.	Valid	1.54
7	*False*	*No*	*No obese can be happy.*	*Raúl is obese.*	*Therefore, Raúl cannot be happy.*	*Valid*	*1.54*
8	False	Ningún	Ningún obeso puede ser feliz.	Raúl no puede ser feliz.	Por tanto, Raúl es obeso.	Invalid	1.54
8	*False*	*No*	*No obese can be happy.*	*Raúl cannot be happy.*	*Therefore, Raúl is obese.*	*Invalid*	*1.54*

For presentation purposes, the materials were distributed in two experimental lists such that a syllogism with a logical order in the first list would have an illogical order in the second list, and vice versa. Nevertheless, the major premises were all the same for the two experimental lists. Participants were randomly assigned to experimental list 1 or 2 and the order in which syllogisms were presented within a list was randomized.

2.3. Experimental procedure

After signing informed consent, participants were fitted with encephalogram (EEG) electrodes while they filled out handedness, vision and health questionnaires. They were seated approximately 100 cm in front of a 19” computer monitor. The session began with a short set of practice stimuli to acclimate the participants to the silently reading and validity decision task. After they read the major premise, the minor premise, and the conclusion, they decided whether the conclusion was logically valid or not. We asked them to do this validity decision to ensure that they would pay attention to the conclusion of the argument. Both initial premises were presented in the screen as a full sentence. The conclusion instead appeared word by word in the center of the screen in order to avoid eye movements and obtain a precise time-lock to the final word of the conclusion. All words in the conclusion were shown in a black 30-point lower-case Arial font on a white background. The major premise was presented in the screen for 3000 ms with an interval of 100 ms before the minor premise. The minor premise

appeared in the screen for 2500 ms. Participants had to press the space bar to initiate the conclusion. Each word of the conclusion was presented for 300 ms with an inter-words interval of 300 ms. Once the conclusion was over, the participants encountered the question: "Do you think the conclusion is logically valid?". They were previously informed that conclusions might be true or false but their task was to decide whether the conclusion correctly followed from the premises. If they thought it was valid, the correct button response was A ("Yes"). If they thought it was invalid, the correct button response was L ("No"). Participants read a total of 160 syllogisms, presented in random order and divided into 3 blocks, with a break between them. Break's duration was unlimited; participants decided when to start the next block. The whole session lasted about thirty-five minutes.

2.4. EEG recording and analyses

EEGs were recorded from 31 tin electrodes mounted in an electrode cap (Electro-Cap International, Eaton, Ohio, USA). Electrode impedances were kept below 5 KΩ. Electrodes were referenced online to the left mastoid, amplified with Brain Amps amplifiers (Brain Products, Munich, Germany) at a sampling rate of 250 Hz with a bandpass of 0.01-40 Hz (electrode sites included: Fp1/z/2, F7/3/z/4/8, FT7/8, FC3/z/4, T7/8, C3/z/4, TP7/8, CP3/z/4, P7/3/z/4/8, O1/z/2, and right mastoid). The electrooculographic activity (EOGs) was recorded using vertical (VEOG) and horizontal (HEOG) bipolar electrodes placed at supra-infraorbital level of the left eye and on the outer canthus of both eyes, respectively. EEG data were analyzed with the Fieldtrip software package (http://www.ru.nl/fcdonders/fieldtrip/), a toolbox implemented in Matlab environment (The MathWorks, Natick, MA). Only trials with a correct response in the validity task were included in the analysis. The continuous sets of raw data were re-referenced to the averaged mastoids and segmented into -100 to 900 ms epochs. An *infomax* independent components analysis (Makeig, Jung, Bell, Ghahremani, & Sejnowski, 1997) was then performed to eliminate the blinks activity (Jung et al., 2000). Finally, epochs

contaminated with gross artefacts were rejected following a z-value visual inspection criteria, a semiautomatic procedure implemented in Fieldtrip. The signal was down pass filtered with a low cut-off at 20 Hz and the activity in the -100 to 900-ms epochs was adjusted to the baseline activity (-100 ms). ERP responses were then assessed using a nonparametric cluster-based random permutation analysis approach (Maris & Oostenveld, 2007). A mass-univariate approach is computed at each spatial and temporal point (Oostenveld et al. 2011). While this approach overrides the problem of a priori choosing locations and/or components, it results in an extremely large number of statistical tests which increase the probability of obtaining false positive results (type 1 error rate). A variety of methods exist to deal with the type 1 error, such as the Bonferroni correction. However, in the context of ERP data, the Bonferroni method is excessively conservative. A popular method to control for the multiple comparison problem is non-parametric statistics. Here we use permutation tests (random shuffles of the data) to obtain the sampling distribution of the test statistic under the null hypothesis. Specifically, permutation tests were used to compute the sampling distribution of a cluster-based statistic. Cluster-based statistics consist in grouping together spatial and temporal adjacent variables (*t* or *F* values for instance) into clusters. The cluster statistic can be defined by its maximal value, extension or a combination of both (Maris and Oostenveld, 2007).Thus, the timing and topographic distributions of logical validity effects for conclusions preceded by true and false major premises was analyzed. The analytic steps were as follows. First, a simple dependent-samples *t*-test for each contrast (invalid vs. valid conclusions from true premises and invalid vs. valid conclusions from false premises) was performed at each time-electrode pair. P-values below 0.05 were used to form clusters of adjacent time points and electrodes. A minimum of two channels were used to form a cluster. Cluster-level test statistic was calculated by taking the sum of all the individual *t*-statistics within that cluster. Then, a null distribution was created by computing 1000 randomized cluster-level test statistics. Finally, the actually observed cluster-level test statistics were compared against the null distribution and

only clusters falling in the highest or lowest 2.5th percentile were considered significant. This procedure allows for the identification of the spatial distribution of validity effects and could effectively handle the multiple-comparisons problem. For significant interactions, we tested our directed hypothesis with respect to a validity

effect (i.e., larger amplitudes for invalid compared to valid conclusions) using one tailed paired t-tests as planned comparisons.

3. Results

Behavioral Results

The behavioral data were analyzed using the statistical program SPSS IBM 22 version (International Business Machines Corp., Armonk, New York, USA). To examine how the type of syllogism, the veracity of the major premise, and the validity of the conclusion influenced the accuracy in the validity task, the number of errors was analyzed for each participant with a mixed factorial design (repeated measures ANOVA) including three within-subject factors (two levels): SYLLOGISM (DARII VS. FERIO), VERACITY (true and false) and VALIDITY (valid and invalid). Whenever the sphericity was violated, we applied the Greenhouse and Geisser (1959) epsilon (ε) correction for degrees of freedom of the within-subject measures. Interaction effects were explored with planned pairwise comparisons. All post-hoc pairwise contrasts were performed and corrected for multiple comparisons by means of the Bonferroni procedure establishing a significance level of $\alpha = 0.05$.

Overall participants were highly accurate in the validity decision task for conclusions following both true (92.5%) and false (91.9%) major premises. The analysis of the number of errors (say logically valid when the syllogism was invalid or say logically invalid when the syllogism was valid) revealed a main effect of Validity [$F(1,28)= 6.86$, $p = 0.014$, $\eta^2_p = 0.19$],

with fewer errors for valid (0.8) than invalid syllogisms (2.2). There was also a main effect of Type of syllogism [$F(1,28) = 9.87$, $p = 0.004$, $\eta^2_p = 0.26$], with a higher number of errors for negative FERIO (1.7) than affirmative DARII syllogisms (1.3). There was also a significant interaction between Veracity and Validity [$F(1,28) = 8.85$, $p = 0.006$, $\eta^2_p = 0.24$]. Post-hoc comparisons revealed significantly more errors for true invalid (5.2) and false invalid (3.8) than for true valid (0.8) syllogisms ($t = -4.22$, $p < 0.001$ and $t= -3.16$, $p=0.004$); and between true (5.2) and false (3.8) invalid syllogisms ($t = 3.97$, $p < 0.001$). Thus, when the syllogisms were invalid, true premises made them more acceptable as valid (more errors) relative to when they followed false premises (fewer errors). The ANOVA revealed as well a significant interaction between the Veracity of the major premise and the Type of syllogism [$F(1,28) = 9.83$, $p= 0.004$, $\eta^2_p = 0.26$]. The comparison between true DARII (affirmative) syllogisms and true FERIO (negative) syllogisms revealed a significant difference ($t=-4.38$, $p < 0.001$) in the number of errors committed. Participants made more mistakes when encountering a true negative (FERIO) syllogisms (3.8) than when encountering a true affirmative (DARIO) syllogisms (2.1). There was also a significant triple interaction between the three factors (Veracity, Validity and Type of Syllogism) [$F(1,28) = 13.62$, $p = 0.001$, $\eta^2_p = 0.32$]. The multiple comparisons showed that most errors were committed for FERIO invalid syllogisms that had followed true premises. They tended to be responded to as logically valid.

Given the high accuracy in the validity task, no participant's ERP data was rejected based on a poor performance. However, individual trials with an incorrect response were rejected for further ERP computations.

3.2. Event-Related Potentials Results

The cluster-based permutation test revealed that the main effect of Veracity was significant [p =.001]. This effect appeared between 350 and 480 ms and was distributed over fronto-central scalp electrodes. Specifically, conclusions derived from true premises had larger negative-going

amplitudes than conclusions derived from false premises. The main effects of Validity and Type of syllogism were not significant [$p > .014$]. However, Validity showed a significant interaction with Veracity [$p < .05$] in the 380-512 ms time-window over centro-posterior electrodes. Planned comparisons showed that, after true premises, invalid conclusions elicited larger negative going responses than valid conclusions [t = -3.3200; p = .0013]. In contrast, after false premises, valid conclusions elicited a larger negativity than invalid conclusions [t = 1.7316; p = .0469] (Figure 1).

3.3. Moral acceptability ratings

Given that the false major premises in our study were often associated with stereotyped social prejudices (e.g. All blond girls are dumb, All Jewish are greedy, All men are sexist) post hoc measures were obtained on the "moral" acceptability of each conclusion. Forty five volunteers, who did not take part in the ERP study, used a 1 to 5-point scale to indicate subjectively whether the conclusion was morally unacceptable (1), morally acceptable (5), or anywhere in between. A 2 by 2 ANOVA revealed a main effect of Veracity [$F(1,79) = 11.26$, $p = 0.001$, $\eta^2_p = 0.125$], with conclusions from true premises more morally acceptable (3.42) than the ones from false premises (3.29). Note that in both cases, conclusions were on average rated in the middle range of morality (3.5 in a 1 to 5-point scale), that is neither moral nor immoral. Critically and in contrast to ERP measurements, the interaction of Veracity and Validity for the judgment of the conclusion´s morality was not significant [$F(1,79) = 0.86$, $p = 0.356$, $\eta^2_p = 0.48$].

Discussion

Our experiment examined the electrical brain activity linked to online syllogistic reasoning with semantic materials. The value of veracity assigned to major premises (true or false) and the logic of the argument (valid or invalid) were manipulated.

For invalid arguments based on true premises and for valid arguments based on false premises, visual inspection suggested that the brain response to the last word of the conclusion elicited a negative-going voltage deflection at around 400 ms (N400). According to cluster-based random permutation analysis, two independent clusters of electrodes became significant at roughly the similar time window (350-480 and 380-512 ms). The first was a fronto-central cluster showing larger negative amplitudes for conclusions after true versus false premises. Its distribution does not correspond to the classical N400 effect and therefore, we will not treat it as an N400 effect. A validity by veracity interaction effect was somewhat delayed (350-475 ms) with respect to common ERP studies of reading for comprehension tasks (Kutas & Federmeier, 2000). Nonetheless, its topographical distribution followed the typical posterior, slightly skewed to the right, maximal amplitude (see Fig. 1B). Interestingly, the direction of this effect varied depending on the interaction of Veracity and Validity (see Fig. 1A).

As expected, when major initial premises were true (e.g. All men are mortal), a larger N400 was elicited for invalid (e.g. Therefore, Juan is a *man*) relative to valid conclusions (e.g. Therefore, Juan is *mortal*). According to recent views on what the N400 indexes (Kutas & Federmeier, 2011), for true premises, participants were able to anticipate the ending of the conclusion thereby reducing the N400 component to valid conclusions. In contrast, participants were troubled by the ending word in an invalid syllogism, eliciting a larger N400 response. Thus, we found a significant effect of validity for categorical syllogisms stemming from true premises, indicating that participants were able to follow the rules of logic to anticipate conclusions. This result was expected based on previous N400 literature. It, however, contrasts with the results obtained by Blanchette & El-Deredy (2014). These authors found increased N400s to any inference making conditions relative to baseline (i.e., the repetition of the minor premise). In addition, our results do not support the hypothesis of the illicit conversion of the premises (Chapman & Chapman, 1959). If conversion had occurred (If

e.g. All men are mortal was taken as the same as All mortals are men), no N400 enhancement would have been present for the invalid conclusion.

Counterintuitively, when initial premises were previously assessed as false (e.g. All blond girls are dumb) a larger N400 was elicited for valid relative to invalid conclusions. Thus, the effect of validity for syllogisms with false premises went in the opposite direction (larger N400s for valid than invalid conclusions) relative to the syllogisms based on true premises. Despite we asked our participants to focus on the logic of the argument by responding whether the conclusion followed logically, their brains reacted with an N400 enhancement to logically valid conclusions drawn from false major premises as if they had not anticipated the "logical" conclusion. In line with one of our hypothesis, word anticipation processes seemed to not be followed whenever the major premise hold untrue. In one case, it was precisely the valid conclusions the ones enhancing the N400 response. In the other case, for invalid conclusions, no N400 was elicited. Therefore, valid yet based on false premises arguments resulted in a difficulty of semantic integration or were not previously anticipated by participants even though the reasoning form was a correct one. In this sense, the interference of beliefs in our logical thinking capacity is supported by our online electrophysiological measures. Previous behavioral studies examined this type of interference. However, they focused on the veracity of the conclusion itself (J. S. T. Evans, J. L. Barston, & P. Pollard, 1983; Revlin et al., 1980). Our study shows that the veracity of the major premise of an argument also influences the response to the processing of the conclusion. This result is difficult to reconcile with the rationality theory (Revlin et al., 1980) that holds that people always follow logical rules. According to this theory, the errors people commit arise by virtue of the wrong encoding of the premises (e.g. illicit conversion). However, when belief in the major premise is manipulated as in our study, beliefs make the processing of valid yet unsound conclusions difficult to process. Thus, logical reasoning is not independent of the veracity of the major premises and their value of truth.

As we mentioned in the introduction, the processing of false statements has previously shown to raise the amplitude of the N400 (ej. Hagoort et al., 2004a). However, our study does not record the response to the word of a sentence that makes a statement untrue. Instead, it presents an initial true or false statement and allows people to keep on reading upstreaming minor premises and their logical/illogical conclusions. The question was whether the correct logic of the argument will make the response to the final conclusion be anticipated even if the initial premise was false (or prejudiced, as we will discuss later). Since the N400 indexes whether upcoming words in discourse could have been anticipated before they appeared on the screen, we were able to determine that logical yet based on false premises conclusions were not anticipated by our participants. The increased amplitude in the N400 time-window for valid versus invalid conclusions of false statements, indicated that they were not anticipated. Therefore, an N400 was still elicited by the highly plausible target word ending "dumb" despite the reasoning threat was a logically correct one. So to say, the inertia to anticipate upcoming words in discourse was overridden by a necessity to believe in the initial major premise. In this regard, our result extends previous results on the brain response to untrue statements themselves (Hagoort et al., 2004a). A perfectly logical and a priori capable of being anticipated ending does however elicit an N400 response in so far as it is preceded by a false or prejudiced statement. On the other hand, when the conclusion was both invalid and based on false premises, an N400 was surprisingly not elicited. We speculate that after reading the second minor premise, participants might have realized that the syllogism did not follow logical rules and they stopped anticipating how it would finish. Future studies are needed to clarify this point.

Finally, it is important to consider that our design included true and false statements that not only differed in veracity rating terms. Some of the statements rated as false by our participants included socially prejudiced statements (e.g. All blond girls are dumb, All Jews are greedy, All men are sexist). This is due to the fact that making false universal categorical

statements (i.e., starting with “All”) often led to such socially prejudiced statements. Consequently, a limitation of our study is a potential confound with the conclusion´s morality. Under those conditions, participants could have elicited an N400 response to the logical conclusion not only grounded on the falsity of the initial premises but also on the “moral” unacceptability of the conclusion (see J. J. A. Van Berkum et al., 2009). However, post hoc analysis on the “moral” acceptability of the conclusionsrevealed that there was no interaction between veracity and validity. Conclusions from false statements, whether valid or invalid, were rated slightly more immoral than the ones from true statements. Note, however, that neither of them was considered to be high in immorality (i.e., which would have corresponded to scores closer to 1 in our morality scale). Moreover, if morality *per se* had an impact on ERP responses, the largest N400 would have been elicited for conclusions from false premises regardless of their validity. In contrast, the N400 results obtained showed no main effects of veracity. Only the interaction between these factors was significant. For true premises, the largest N400 was elicited by invalid syllogisms whereas for false premises, the largest N400 was elicited for valid syllogisms.

Our study suggests that among the theories of syllogistic reasoning (Khemlani & Johnson-Laird, 2012), the mental model theory of reasoning (Johnson-Laird, 1975) is the best to fit the ERP pattern of results. This theory posits than individuals grasp that an inference is no good if there is a counterexample to it (cited in Khemlani & Johnson-Laird, 2012). Participants in our study potentially found counterexamples to the false major premises (e.g. intelligent blond girls) and, as a consequence, they had difficulties to process a conclusion that despite being drawn on perfectly logical grounds disregarded veracity, relative to the condition in which the initial major premise was considered both valid and true.

With regard to the type of syllogism, behavioral errors in the validity judgment task significantly increased for FERIO relative to DARII syllogisms. However, whether affirmative or

negative propositions were used (DARII and FERIO syllogisms, respectively) had no main effect, nor interactions with other factors in terms of ERP response.

In conclusion, human brains are rational: they do react to invalid conclusions drawn from true premises. However, in line with previous ERP studies (Hagoort, Hald, Bastiaansen, & Petersson, 2004b; J. J. Van Berkum, B. Holleman, M. Nieuwland, M. Otten, & J. Murre, 2009) human brains are also knowledge-biased during online categorical thinking: we cannot conceive of a valid argument if it clashes with our previous world knowledge and beliefs. The ERP technique is a useful tool to explore online reasoning with semantic materials. Our study reveals how prior beliefs, mostly related to socially unacceptable statements, override the anticipation of logical conclusions, in particular the processing of valid conclusions that are, however, grounded on false major premises.

References

Blanchette, I., & El-Deredy, W. (2014). An ERP investigation of conditional reasoning with emotional and neutral contents. *Brain Cogn, 91*, 45-53. doi:10.1016/j.bandc.2014.08.001

S0278-2626(14)00127-4 [pii]

Bonnefond, M., & Henst, J. B. (2013). Deduction electrified: ERPs elicited by the processing of words in conditional arguments. *Brain Lang, 124*(3), 244-256. doi:10.1016/j.bandl.2012.12.011

S0093-934X(13)00007-2 [pii]

Bonnefond, M., Kaliuzhna, M., Van der Henst, J. B., & De Neys, W. (2014). Disabling conditional inferences: an EEG study. *Neuropsychologia, 56*, 255-262. doi:10.1016/j.neuropsychologia.2014.01.022

S0028-3932(14)00040-2 [pii]

Bonnefond, M., & Van der Henst, J. B. (2009). What's behind an inference? An EEG study with conditional arguments. *Neuropsychologia, 47*(14), 3125-3133. doi:10.1016/j.neuropsychologia.2009.07.014

S0028-3932(09)00303-0 [pii]

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of experimental psychology, 58*, 220-226. Retrieved from <Go to ISI>://MEDLINE:13809244

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat Neurosci, 8*(8), 1117-1121. doi:10.1038/nn1504

Evans, J. S., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem Cognit, 11*(3), 295-306. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6621345

Evans, J. S. B. T., Thompson, V. A., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in psychology, 6*, 398. Retrieved from <Go to ISI>://MEDLINE:25904888

Evans, J. S. T., Barston, J. L., & Pollard, P. (1983). On the Conflict between Logic and Belief in Syllogistic Reasoning. *Memory & Cognition, 11*(3), 295-306. doi:Doi 10.3758/Bf03196976

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491-505. doi:PSYP531 [pii]

10.1111/j.1469-8986.2007.00531.x

Greenhouse, S. W., & Geisser, S. (1959). On Methods in the Analysis of Profile Data. *Psychometrika, 24*(2), 95-112. doi:Doi 10.1007/Bf02289823

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004a). Integration of word meaning and world knowledge in language comprehension. *Science (New York, N Y), 304*(5669), 438-441. doi:10.1126/science.1095455

1095455 [pii]

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004b). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438-441. doi:10.1126/science.1095455

Johnson-Laird, P. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 7-54). Springdale, NJ: Erlbaum.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology, 37*(2), 163-178. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10731767

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychol Bull, 138*(3), 427-457. doi:10.1037/a0026841

2012-02603-001 [pii]

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn Sci, 4*(12), 463-470. doi:S1364-6613(00)01560-6 [pii]

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu Rev Psychol, 62*, 621-647. doi:10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science (New York, N Y), 207*(4427), 203-205. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7350657

Makeig, S., Jung, T. P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proc Natl Acad Sci U S A, 94*(20), 10979-10984. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9380745

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods, 164*(1), 177-190. doi:10.1016/j.jneumeth.2007.03.024

Oaksford, M., & Chater, N. (2009). Precis of bayesian rationality: The probabilistic approach to human reasoning. *Behav Brain Sci, 32*(1), 69-84; discussion 85-120. doi:10.1017/S0140525X09000284

Petten, C. V., Kutas, M., Kluender, R., Mitchiner, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *J Cogn Neurosci, 3*(2), 131-150. doi:10.1162/jocn.1991.3.2.131

Prado, J., Chadha, A., & Booth, J. R. (2011). The brain network for deductive reasoning: a quantitative meta-analysis of 28 neuroimaging studies. *J Cogn Neurosci, 23*(11), 3483-3497. doi:10.1162/jocn_a_00063

Qiu, J., Li, H., Huang, X., Zhang, F., Chen, A., Luo, Y., . . . Yuan, H. (2007). The neural basis of conditional reasoning: an event-related potential study. *Neuropsychologia, 45*(7), 1533-1539. Retrieved from <Go to ISI>://MEDLINE:17194466

Revlin, R., Leirer, V., Yopp, H., & Yopp, R. (1980). The belief-bias effect in formal reasoning: the influence of knowledge on logic. *Mem Cognit, 8*(6), 584-592. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7219179

Striker, G. (2009). *Aristotle´s Prior Analytics. Book 1.*: Oxford University Press.

Van Berkum, J. J., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychol Sci, 20*(9), 1092-1099. doi:10.1111/j.1467-9280.2009.02411.x

Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or Wrong? The Brain's Fast Response to Morally Objectionable Statements. *Psychological Science, 20*(9), 1092-1099. Retrieved from <Go to ISI>://000269391900010

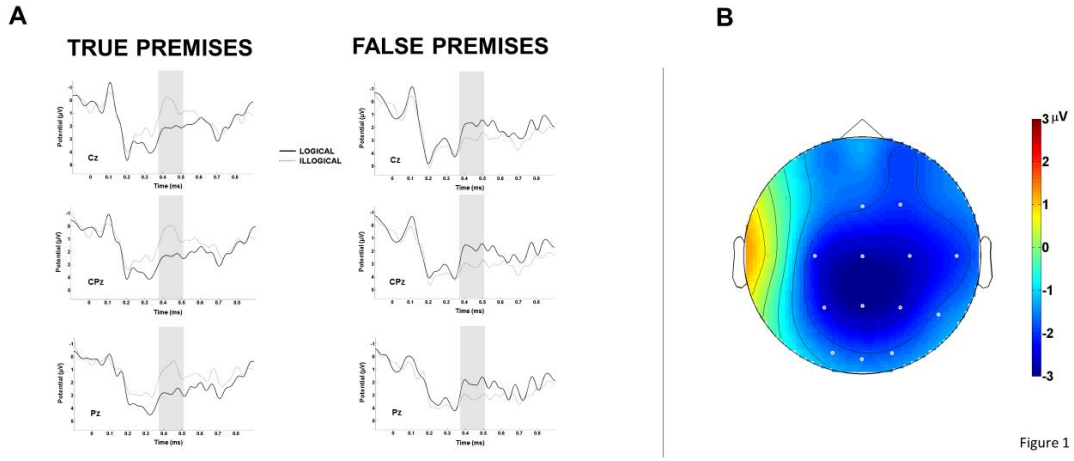

Figure 1

Figure caption

Figure 1. A. Grand averaged ERP response to the final conclusion of syllogisms when true (left) or false (right) major premises preceded at three representative electrodes (Cz, CPz, Pz) Solid lines represent the response to logically valid conclusions. Light grey lines depict the response to logically invalid conclusions. According to cluster-based computation analysis, the shaded

grey area indicates where conditions statistically differed, in the 380-512ms time-window. **B.** Topographical map of voltage. The ERP response to valid conclusions was subtracted from the one to invalid conclusions. The plot represents the difference between validity effects in true and false conditions. The set of electrodes included in the significant cluster are marked as white circles. A typical N400 centro-parietal distribution is observed.

Funding

This work was supported by the *Ministerio de Economía y Competitividad* (MINECO) [grant numbers: PSI2014-60682 to EMM and PSI2015-68368-P-MINECO-FEDER to JAH], and by the *Comunidad Autónoma de Madrid* [grant number H2015/HUM-3327 to EMM and JAH].

Acknowledgements

We thank all the study participants for their cooperation and Pilar Lizano for her assistance in data collection.