When Does a Reasoner Respond: Nothing Follows?

Hannah Dames
University of Zurich

Daniel Brand
University of Freiburg

Nicolas Riesterer
University of Freiburg

When Does a Reasoner Respond: Nothing Follows?

Marco Ragni* (ragni@cs.uni-freiburg.de), Hannah Dames*(damesh@cs.uni-freiburg.de), Daniel Brand (daniel.brand@cognition.uni-freiburg.de), Nicolas Riesterer (riestern@cs.uni-freiburg.de),
Cognitive Computation Lab, Department of Computer Science, Georges-Köhler-Allee 52, 79110 Freiburg, Germany

Abstract

When does a reasoner respond that ”no valid conclusion” (NVC) follows in a syllogistic reasoning task? Cognitive theories aim to trace it back to theory specific inference processes. In contrast, systemic theories explain it by depleted cognitive resources among others. This paper investigates possible theories to explain NVC responses in an experiment with 139 participants. Using mixed models we analyze the association of NVC responses with reaction times, the validity as well as the entropy of a syllogism, and how NVC responses change over time. As expected, the number of NVC responses is lower than logically expected, participants respond NVC more often for invalid syllogisms, and the likelihood to respond NVC increases over the time-course of the experiment. Surprisingly, however, only for valid syllogisms, are the entropy and the RTs associated with NVC responses. Consequently, for invalid syllogisms, NVC responses seem to be generated differently as compared to valid ones.

Keywords: Reasoning; NVC; cognitive theories; logic; valid; invalid

Introduction

The psychology of reasoning investigates when and which conclusion is derived from given information. This includes the case when *no conclusion* can be drawn because the information is insufficient or it is too difficult to make an inference. The domain of syllogistic reasoning is probably the best researched domain with most published theories (for an overview see Khemlani & Johnson-Laird, 2012). A *syllogism* consists of two quantified statements. Each statement is formed using one of four quantifiers: All (A), Some (I), Some ... not (O), or None (E). Consider the following syllogistic reasoning problem:

(AA4) All beekeepers are architects.
All beekeepers are chemists.
———————————————
What, if anything, follows?

The task is to generate a quantified answer using one of the quantifiers A, I, O, E about the two terms architects (A) and chemists (C, in any direction) or to conclude that no logically valid conclusion (*NVC* for short) can be made. Four different arrangements of the terms in the premises, called figures, are possible. The example above, for instance, is a type 4 figure (B-A, B-C). The four quantifiers for each of the two premises times four figures sum up to 64 possible syllogistic problems. Each syllogism can be encoded by a string, describing the quantifiers of the two premises as well as the relation of the used items in a figure. Hence, the syllogism above can be succinctly written as AA4. For the problem (AA4) most of the participants (49% in Khemlani & Johnson-Laird, 2012) infer that *All architects are chemists* and only 16% give one of the logical correct answer that *Some architects are chemists* or that *Some chemists are architects*. However, about 22% of the participants in the metaanalysis (Khemlani & Johnson-Laird, 2012) respond that NVC follows. A *logically valid* problem is one where by applying a logical calculus such as first-order logic allows to infer a conclusion (such as the syllogistic example AA4 above where *Some architects are beekeepers* is one). If this cannot be inferred then it is called *invalid* problem (and the only logical correct answer is NVC). Past research both from a statistical and from a modeling perspective has strongly focused on the case when an inference can be drawn (Oaksford & Chater, 2007; Johnson-Laird, 2006; Costa, Saldanha, Hölldobler, & Ragni, 2017) but less on the case when *no logically valid conclusion* can be inferred. Yet, it is exactly this response that stands out from the rest: Not only is the response NVC a *different class of response*, namely stating that no other conclusion follows, but it is the NVC response, that is the *most frequently observed response* in experiments (Khemlani & Johnson-Laird, 2012). In the current work, we aim to fill the gap of investigations on NVC responses by systematically investigating *when* people respond NVC. In particular, we compare different approaches to explain NVC responses by analyzing experimental data.

When is an NVC response given?

Syllogistic theories have been categorized as heuristic, rule-based, and model-based approaches (Khemlani & Johnson-Laird, 2012): Only few cognitive theories in syllogistic reasoning predict the NVC conclusion at all (e.g., Mental Model Theory, Verbal, Conversion). If a theory does so, it often implies that individuals give NVC as a last-resort, when the inference process yields nothing else (e.g., Mental Logic; Rips, 1994). Most of the heuristic theories do not predict NVC responses, with a rare exception in the case of Conversion and the probabilistic heuristic model (PHM, Oaksford & Chater, 2007, but see Copeland, 2006) that can be extended to predict NVC. The Atmosphere (Woodworth & Sells, 1935) and Matching (Wetherick & Gilhooly, 1995) theories derive only the quantifier in the response from the premise quantifiers. Hence, they do not consider and cannot explain NVC responses. This is remarkable as in the case of syllogisms there are 37 invalid problems (58% of all syllogisms) that would require from a normative logical perspective NVC as the correct response.

In sum, while there are at least twelve cognitive theories about syllogistic reasoning (Khemlani & Johnson-Laird, 2012), there is no explicit cognitive reasoning theory beyond explaining it by a search through the theory specific inference mechanism (e.g., by applying all inference rules or the generation of all models). Beyond explaining NVC by *cognitive reasoning theories*, *systemic theories* can provide alternative accounts emphasizing the role of behavioral response tendencies within experiments. Among others these systemic hypotheses include phenomena such as mental depletion (i.e., with each syllogism the cognitive resources are depleted (e.g., Schmeichel & Vohs, 2018) or cognitive load

*Both authors contributed equally to this manuscript.

(e.g., Sweller, 1994), which lead an individual to stop reasoning as soon as the problem becomes too difficult, or a general aversion to respond NVC (e.g., NVC is interpreted as ”giving-up”). These hypotheses including their assumed, underlying cognitive processes are briefly summarized in Table 1. The aim of the current paper is to investigate such cognitive and systemic hypotheses in explaining why a NVC response is given. For the above-mentioned differences in the ability to predict NVC within heuristic theories, we focus on the MMT and mental logic theory in the current paper. Thereby, we aim to provide novel and much needed insights into when participants respond *no valid conclusion*.

Theories, predictions, and hypotheses

We have identified distinctive cognitive theories and systemic hypotheses that can explain when a reasoner responds NVC. In this section, we briefly outline these theories as well as hypotheses and draw implications on five observable patterns of NVC: first, the *response time* (RT) and the *frequency* of NVC response. Furthermore, we investigate the influence of valid and invalid syllogisms (*validity*) and the problem's entropy. The entropy measure (Shannon & Weaver, 1963) has been applied to measure the response diversity of each syllogism (Khemlani & Johnson-Laird, 2012). For each syllogism per study, the authors computed the probability with which each conclusion was drawn and aggregated the probabilities using Shannon's measure. The response diversity demonstrates an uncertainty of reasoners about which conclusion has to be drawn. Last, we analyze how the likelihood to respond NVC changes over the course of 64 syllogisms.

Theory of Mental Models (MMT). The MMT (e.g., Johnson-Laird, 2006) postulates a two-stage process based on the generation of an initial model and a flesh-out process that tests a putative conclusion formed on the initial model by a search for counter-examples (e.g., Bucciarelli & Johnson-Laird, 1999). If the flesh-out process does not yield a conclusion, the reasoner responds NVC. The latest implementation, mReasoner[1], contains a specific parameter that guides the generation of counter-examples. MMT makes the following predictions. *RT of NVC*: On average, the MMT predicts the NVC response at the end of the inference process, hence, responding NVC requires more cognitive steps and thus more time (especially, in the case of multiple model problems, i.e., problems that are invalid). *Frequency of NVC*: The inference process described before, however, can sometimes fail or be stopped early. As a result, not in all cases counter-examples are searched for and putative conclusions are drawn, even in cases where NVC hold. Consequently, less NVC responses are given as required by formal logic. *Entropy and NVC*: In indeterminate cases, the flesh-out process becomes relevant, hence, the more difficult a problem is or the more uncertainty it causes (measured by the entropy), the more NVC responses will be generated. *Time-course of NVC*: The more syllogisms are solved, participants enter more likely the flesh-out process (reasoners become more logical, as it has been recently modeled in mReasoner; Ragni, Riesterer, Khemlani, & Johnson-Laird, 2018). Consequently, participants are more likely to respond NVC for invalid syllogisms over time.

[1] `https://mentalmodels.princeton.edu/models/mreasoner/`

Theory of Mental Logic (ML). The theory of ML (Rips, 1994) is based on the application of first-order formal inference rules together with the inclusion of Gricean implicature to capture differences between a formal and an everyday language understanding of existential quantifiers. As it is based on formal logic rules, the conclusions are valid and no erroneous results will be predicted. The theory proposes that the erroneous responses generated by human reasoners are due to problems in the recognition, retrieval, or application of the formal rules (Rips, 1994). Following predictions can be derived: *RT of NVC*: ML predicts NVC, if the full application of the inference mechanism does not yield a conclusion. This takes longer than the application of some inference rule in the valid case. *Frequency of NVC*: An NVC response is found in the invalid cases and not in the valid cases. *Entropy and NVC*: A connection has not been reported and so we do not assume a predicted difference. *Time-course of NVC*: The mental logic does not assume a change across time.

Predictions of Mental Depletion. Theories of resource depletion (e.g., Schmeichel & Vohs, 2018) assume that mental activities such as reasoning can deplete cognitive resources. This results in an increase in NVC responses over time due to depletion. This increase appears for valid and invalid syllogisms - due to the depleted cognitive resources. A simple depletion model makes no distinction between logically valid and invalid problems. While combinations with cognitive theory can be thought of, we solely focus on the case where more NVC responses are given over time. Predictions: *RT of NVC*: No effect of NVC-responses on RTs is expected. Depletion processes may result in either generally higher or lower RTs over the course of an experiment, but regardless of an NVC response. *Frequency of NVC*: There are no concrete predictions. *Entropy and NVC*: Entropy has no implications on NVC, but instead generally enhance mental depletion. *Time-course of NVC*: Mental depletion is assumed to strengthen throughout an experiment. Thus, NVC responses should increase across the course of solving the 64 problems, respectively.

Predictions of Early Stoppers. Some syllogisms are more difficult than others and thus require additional cognitive resources. For some syllogisms, reasoners may stop the reasoning process early avoiding the mental effort required by analytic processes by responding NVC. While the application of heuristics would not result in an NVC answer, the early stopping process does (NVC as a last resort). Early stoppers do not necessarily make a distinction between valid and invalid problems as for both types problems with a high entropy exists. Following predictions are derived: *RT of NVC*: An early stopper does not need longer for an NVC response. *Frequency of NVC*: Both valid and invalid problems can be difficult to solve. Therefore, the early stopper hypothesis predicts generally more NVCs as there are logically correct NVC responses. *Entropy and NVC*: Higher entropy resembles a higher uncertainty with the problem at hand, which may lead to more NVC responses the higher the entropy. *Time-course of NVC*: The time-course has no effect.

Predictions of NVC aversion. Logically naive reasoners may interpret responding NVC as “giving up” (similar to the last-resort option as it is assumed in many theories). While participants may even fear to be regarded as less intelligent or ignorant, they may (at least in the beginning) tend to avoid this answer. The following predictions can be made: *RT of NVC*: NVC aversion leads to higher RTs for NVC responses as the deliberation processes to exclude all other response is time-consuming. *Frequency of NVC*: As NVC is avoided, fewer NVC responses as there are logically correct ones are made. *Entropy and NVC*: It is unclear whether Entropy may have an effect. *Time-course of NVC*: The aversion for NVC may diminish over time due to exposition to invalid syllogisms or because the reasoner learns that some syllogisms do not have a valid response. Hence, NVC responses increase over time.

Hypotheses

The introduced theories and hypotheses differ on predictions for response times, frequency of NVC answers, entropy, and the time-course of NVC. Based on these predictions, we will derive five general hypotheses. The presented cognitive theories and hypotheses do explain an NVC response in one of two ways: by the application of the complete inference mechanism that does not yield any valid conclusion or by a model-based search that yields counter-examples to any putative valid conclusion. This implies, however, that more steps are necessary to infer that nothing follows than to infer that something follows. More cognitive steps, however, require more time. This leads to our first hypothesis: *Hypothesis 1: The RTs significantly increase in trials where a NVC response is given as compared to non-NVC trials.*

Cognitive reasoning theories assume that NVC is a response typically generated after the application of inference rules or through the search through all counter-examples. This process is not necessarily always entered resulting in the miss of NVC responses. Thus: *Hypothesis 2: The number of NVC responses is lower than the number of logically correct NVC responses.*

Since validity is a logical concept, cognitive theories that are closer to logic make a difference between them. Hence, we get as a corollary hypothesis: *Hypothesis 3: The number of NVC responses is lower in the case of valid problems than in the case of invalid problems.*

Moreover, if it is more likely for a reasoner to respond NVC, if there is greater uncertainty operationalized by entropy. *Hypothesis 4: The higher the entropy of a syllogism the higher the likelihood of an NVC response.*

A fifth hypothesis is that across an experiment participants may increasingly respond NVC, which can depend both on cognitive (e.g., MMT) and systemic hypotheses (e.g., mental depletion): *Hypothesis 5: There is an increase in NVC responses across solving more problems.*

The different predictions of the cognitive reasoning theories and systemic hypotheses are summarized in Table 1. In the next section we report experimental data and the analysis.

Table 1: The hypotheses and predictions of the cognitive theories and the systemic factors.

Theories	**Prediction**				
	RT H1	NVC H2	Validity H3	Entropy H4	Time H5
Mental Model	y	y	y	y	y
Mental Logic	y	n	y	n	n
Mental Depletion	n	?	n	n	y
Early Stopper	n	n	n	y	n
NVC aversive	y	y	n	?	y

Explanation of the abbreviation y = the theory predicts yes; ? = the theory does neither predict yes nor no; n = the theory predicts no.

Experiment

Method

The experiment tested 204 participants (125 female and 79 male) on Amazon's Mechanical Turk[2]. They received a nominal fee for their participation. Participants or trials were excluded based on the following criteria: First, in order to identify non-compliers, data from participants that are at or below guessing level were discarded. The cutoff point of 18.8% (n = 64) is calculated as the cumulative binomial probabilities of 1/9 (for 9 possible conclusions) for 64 correct responses. That results in twelve problems correct for the α-value of .05 according to the binomial distribution. Second, trials with exceptionally long response times (RT) were excluded from the analyses: RTs exceeding 10 minutes (n = 1) and RTs deviating more than 3 standard deviations (SDs) from the individual mean RT separated for valid vs invalid syllogisms (n = 147, 1.7% of remaining trials). Last, the first four trials of the experiment were excluded as the four first trials always consisted of the same syllogisms for practice purposes (n = 546). Thus, 139 participants and 8202 observations were included in the following analyses.

Each participant had to select a conclusion from all possible nine response options for all 64 syllogisms (selection task). The order of the problems was randomized for each participant, except that the problems, AA1, AI2, EA3 and IA4, were always presented first in a randomized order, so that participants can familiarize themselves with the experiment. In addition, four single-premise syllogisms (of the four different possible quantifiers) were used as practice trials. Participants received two assertions similar to problem AA4 above. Content was randomly assigned to all 64 syllogisms (thus, valid and invalid problems received the same content with similar premise lengths of the resulting premises). For each set they had to determine which eight possible conclusions logically follow from the assertions by pressing one of the eight keys: 1-4 (the respective quantifier with the conclusion direction A-C) and 7-0 (the respective quantifer with the conclusion direction C-A). If no logical conclusion could be found, participants had to press the space bar. There were eight presentation orders of the conclusion quantifiers to reduce the presentation order effect. Each participant received the same response option order throughout the whole experiment. They could take as much time as they needed, but responses within a second were prohibited.

[2]https://www.mturk.com

Results

The overall percentage of logically correct responses per participant was 38.7% (*SD* = 19.0%), for the 27 syllogisms with valid conclusion(s) 42.1% (*SD* = 15.3%) and for the 37 syllogisms without a valid conclusion (NVC syllogisms) 36.5% (*SD* = 27.1%). On average, for valid syllogisms, participants gave 16.9% NVC responses (*SD* = 16.6%) and 36.5% (*SD* = 27.1%) for invalid syllogisms.

Analysis. Participants' frequency of NVC-responses differed between individuals (*M* = 29.0%, *SD* = 21.4%). In fact, there were a few participants that did not give any (*n* = 8) or less than 10 (*n* = 48) NVC responses. In the following analyses we used (generalized) linear mixed models (short (G)LMM; for an overview see Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012) as they can handle incomplete and unbalanced data and can account for the multi-level structure of the designs (e.g., multiple measures per participant). GLMMs were analyzed using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015, Version 1.1.19) in the R environment. Models were fit via maximum likelihood (ML). Effect coding was used for all dichotomous fixed effects. Denominator degrees of freedom and *p*-values were estimated via Satterthwaite corrections implemented via lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017, Version 3.0.1). Furthermore, the significance of fixed effect on the model fit was obtained by step-wise removing a fixed effect from the full model and testing whether the exclusion of the variable resulted in a significant loss of the goodness of fit as indicated by likelihood ratio tests and by comparing the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). The reported tables (Table 2 and 3) show the results for the best models.

The analysis of reaction times. For the analysis of RTs, there is currently a debate about whether or not dependent variables should be transformed (Lo & Andrews, 2015). It has been suggested to use GLMMs on the raw RTs to analyze non-normal data that involve random effects (Lo & Andrews, 2015). Here, were use Inverse Gaussian distributions to account for the distinct positive skewed distribution of the continuous, raw RTs (for an overview of this approach, see Lo & Andrews, 2015). However, this approach resulted in a significantly worse fit (χ^2 = 156020, $p < .001$) than the standard logarithmic approach using LMMs (where RTs were logarithmically transformed prior to analyses). As we report only the best-fit models, we therefore only display the LMMs on the logarithmically transformed RTs. However, results were similar both in the transformed and the untransformed analysis. The RTs were analyzed using LMMs with the factors *validity* (invalid = -1 vs. valid = 1), the *"NVC"* response (No NVC = -1, NVC = 1), and the corresponding interaction as fixed factors (1). We implemented the maximal random-effects structure justified by the design (as suggested by Barr, Levy, Scheepers, & Tily, 2013): Participants (including by-participant random slopes for Validity, NVC, and their interaction) and the different syllogism problems were treated as a random factors (2). The trial *"sequence"* (4-64) was added as covariate since it correlated with the NVC response capturing effects due to fatigue or learning (1). All continuous predictor variables were centered and scaled. The full model was specified as follows:

$$log(RT) = NVC * Validity + Sequence \quad (1)$$
$$+ (NVC * Validity | Participant) + (1 | Syllogism) \quad (2)$$

The results of the best-fit model can be taken from Table 2.

Table 2: Fixed-Effect Parameter Statistics for the full/ best-fit Reaction Time model.

Predictors	Estimates	*SE*	*t*	*p*
Intercept	9.43	0.05	176.76	$< .001$
NVC (yes = 1)	-0.02	0.02	-1.12	*.270*
Validity(valid = 1)	0.06	0.02	3.51	*.001*
Sequence	-0.13	0.01	-22.30	$< .001$
NVC:Validity	0.05	0.01	4.80	$< .001$

Hypothesis 1: Other than expected, there was no main effect of NVC on the RT as the RTs did not significantly increase in trials where a NVC response was given as compared to non-NVC trials. However, there was a significant interaction between NVC responses and the validity of the syllogism: the RTs were significantly associated with the occurrence of a NVC responses for valid syllogisms. In trials with NVC responses, the RT increased, but only for valid syllogisms. Any reduction of a parameter (e.g., of the interaction) resulted in a significantly worse model fit as compared to the full model reported. The interaction was also apparent in the mean RTs: For valid syllogisms, the RTs were higher for NVC responses (*M* = 20.17, *SD* = 17.37) as compared to other conclusions (*M* = 16.55, *SD* = 7.64). However, there was no difference for invalid syllogisms (NVC: *M* = 16.42, *SD* = 11.26, Other conclusions: *M* = 16.56, *SD* = 8.30).

The analysis of the likelihood to give a NVC responses. The occurrence of NVC-responses as a bivariate dependent variable was analyzed using GLMMs (NVC response = 1, no NVC response = 0). GLMM estimates were computed with a logit link, binomially distributed residuals using the "bobyqa" optimizer with 200 000 iterations. Odds ratios (ORs) of the fixed effects coefficients of the full model are reported as effect sizes.

The occurrence of a NVC response was analyzed with the factors Validity (invalid = -1 vs. valid = 1), the Entropy of each syllogism (using the entropy measures computed by Khemlani & Johnson-Laird, 2012), as well as the corresponding interaction and the trial sequence (4-64) as fixed factors (3). We again implemented the maximal random-effects structure justified by the design: Participants (including by-participant random slopes for the factors Validity, NVC, and their interaction) and the syllogism problems (random intercept) were treated as a random factors (4). The entropy variable was centered prior to analysis. The full model of was specified as follows:

$$NVC = Validity * Entropy + Sequence \quad (3)$$
$$+ (Validity * Entropy | Participant) + (1 | Syllogism) \quad (4)$$

Table 3: Fixed-Effect Parameter Statistics for the best-fit NVC model.

Predictors	Estimates	*SE*	*z*	*OR*	*p*
Intercept	1.77	0.19	-9.23	0.17	$< .001$
Validity$_{(\text{valid}=1)}$	-0.77	0.15	-5.20	0.46	$< .001$
Entropy	0.41	0.35	1.15	1.5	*.249*
Sequence	0.18	0.03	5.62	1.19	$< .001$
Validity:Entropy	1.08	0.35	3.05	2.94	$= .002$

Note. OR indicates Odds Ratios.

The results of the best-fit model can be taken from Table 3.

Hypothesis 2. In 53% of the syllogisms a NVC response is the logically conclusion. As hypothesized, in the current experiment, participants gave 28.99% NVC responses (*SD* = 21.40%) on average which is significantly less than 58% ($V = 382$, $p < .001$; a paired Wilcoxon signed tank test was used due to a deviation from normality). Thus, we can confirm that the number of NVC responses was lower than the number of logically correct invalid syllogisms.

Hypothesis 3. As hypothesized, the occurrence of a NVC response was significantly associated with the validity of the syllogism. NVC responses were more likely to occur for invalid than for valid syllogisms. Excluding this factor from the full model resulted in a significant reduction of the overall fit ($\chi^2 = 189.56$, $p < .001$).

Hypothesis 4. We expected that the higher the entropy of a syllogism was the higher the likelihood of a NVC response would be. Other than hypothesized, there was no significant main effect for entropy on the likelihood to give a NVC response. However, there was a significant interaction between validity and entropy (see Figure 1 for an illustration): Entropy impacted the likelihood to respond NVC, but only for valid syllogisms. Excluding this interaction as well as the entropy factor from the full model resulted in a significant reduction of the overall fit ($\chi^2 = 57.07$, $p < .001$). A post-hoc analysis for the number of NVC responses and entropy also revealed a strong association between entropy and the relative frequency of NVC responses for each syllogisms for valid ($r_\rho = .69$, $p < .001$) but not for invalid syllogisms ($r_\rho = -.27$, $p = .112$).

Hypothesis 5. The effect of the trial sequence on the relative frequency of NVC responses separated for valid and invalid syllogisms is illustrated in Figure 2. The plot highlights that NVC responses do not stay constant over the time-course of the experiment. As expected, there was also a significant main effect of the sequence on the likelihood to give a NVC response in the mixed model (see 3). Excluding the sequence factor from the full model resulted in a significant reduction of the overall fit ($\chi^2 = 30.63$, $p < .001$). Since NVC is a logically sound conclusion only for invalid syllogisms, a simultaneous increase for invalid and decrease for valid syllogisms would indicate a trend towards a support for the theory that reasoners become more logical in the experiment. However, the increase in NVC response probability does not differentiate between valid and invalid syllogisms.

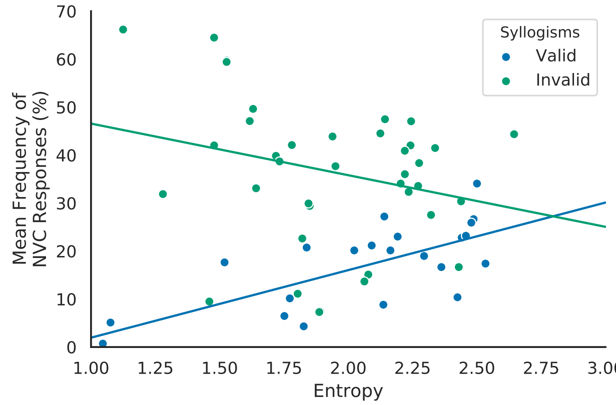

Figure 1: The relationship between the frequency of NVC responses and entropy. Linear regression lines are plotted separately for valid and invalid syllogisms.

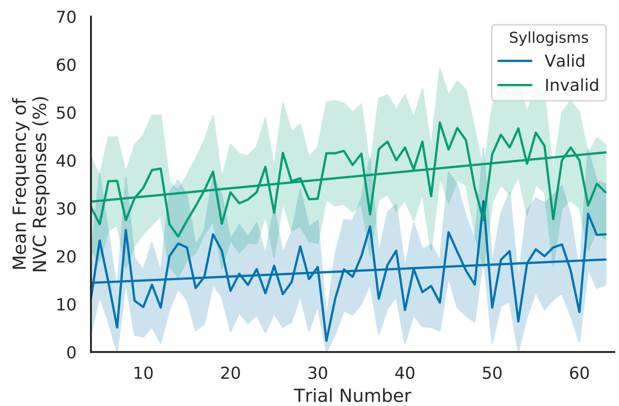

Figure 2: Mean number of NVC responses on valid and invalid syllogisms for the experimental sequence of 64 syllogisms.

An upwards trend can be observed for both. Furthermore, in follow-up analyses the inclusion of an interaction effect for the sequence and the validity of a syllogisms did not result in a fit improvement to the reported best-fit model $\chi^2 = 0.65$, $p = .420$.

Discussion

On which factors does the likelihood to respond NVC depend and how are NVC responses associated with differences in RTs? First, the RTs seem to increase in NVC trials as compared to trials where another conclusion was given - but only (and other than expected) for valid syllogisms. There are various explanations why RTs did not increase for NVC responses in invalid syllogisms. For instance, NVC responses are the "logical" correct response option for invalid syllogisms. Thus, on average, NVC responses for these problems could occur for both logical reasoning and as a consequence of other processes (e.g., guessing, giving-up, etc.). As a consequence, responding NVC may not only be a "last resort" after elaborate reasoning (thus, higher RTs), but also stem from logically correct reasoning. Also, providing any response other than NVC for invalid syllogisms is logically incorrect and may therefore include deviating processes (possibly leading to prolonged RTs). Therefore, for invalid syllogisms, such effects may mask the effect of NVC responses. Second, while we found a significant main effect of validity on NVC, entropy was associated with NVC responses only for valid syllogisms. The reported interaction between validity and entropy on the frequency of NVC responses is however only logical. As theorized, the frequency of NVC responses seems to be higher for high entropy problems as compared to low entropy problems for valid syllogisms. The opposite relationship observed for invalid problems is logical as naturally for easy invalid syllogisms (reflected in a low entropy), participants should most frequently respond "NVC". The harder an invalid syllogism becomes (possibly reflected by a high entropy), the more the responses spread, and the less often a NVC response is given. Future analyses should investigate whether NVC responses are selected more frequently for high entropy problems in addition to the general benefit or drawback NVC responses receive by a higher variance in responses. Third, as hypothesized, the likelihood to respond NVC increases with the trial sequences during the time-course of the experiment. Surprisingly, this association seems to be apparent for both valid and invalid syllogisms. The effect of trial sequence on NVC responses can thus not be explained by participants becoming more logical. On the contrary, the results point towards other systemic processes taking place during the course of the experiment. Note, that this study used a selection task. It remains an open question how our results relate to tentative studies on generation tasks (for an overview of differences of response formats see Hardman & Payne, 1995). Moreover, variations of the classic syllogism task, such as the countermodel/ "Harry"-task (see Achourioti, Fugard, & Stenning, 2014), would certainly provide additional insights on the questions when participants conclude that "nothing follows" in other test situations.

General Discussion

When does a reasoner respond "nothing follows"? To answer that question we have investigated implications of the mental model (e.g., Johnson-Laird, 2006) and mental logic theory (Rips, 1994) as well as adapted alternative systemic hypothesizes such as the role of mental depletion. First, reasoners seem to take longer when responding NVC only for valid and not for invalid syllogisms. With regard to the proposed theories and systemic hypotheses of interest, this finding poses a challenging novel perspective on NVC responses as this distinction is not yet predicted by cognitive theories: giving a NVC response generally takes longer due to the requirement of more cognitive steps, e.g., by generating all inferences or searching for counterexamples (Khemlani & Johnson-Laird, 2012). So, the time needed to respond "nothing follows" is expected to be independent of the validity of a problem. Moreover, the Early Stopper hypothesis contradicts this empirical finding: An Early Stopper would not need more time for responding NVC. In sum, our assumptions holds true only for valid syllogisms. This raises the question whether invalid and valid syllogisms are processed differently and influenced by other processes such as mental depletion or a NVC aversion. Second, the likelihood to respond NVC increases for both valid and invalid syllogisms over time indicating that these differences cannot be explained by participants becoming more logical within the same experiment. While the Early Stopper hypothesis cannot account for this finding, the results can be well explained by the NVC aversion hypothesis. Participants may have an early aversion to respond NVC. If the NVC response is assigned a meaning of "I give up", participants might need to encounter some of the invalid syllogisms to gain confidence in stating that no conclusion may follow from the premises. It is possible that a reasoner may for instance learn across solving syllogistic problems that for some types of problems a valid conclusion cannot be found. Hence, the reasoner can start to assume that the probability of NVC problems is high (with each such observation). The aversion may however also diminish over time due to depletion or fatigue effects.

What can we conclude regarding our proposed theories and systemic hypotheses based on these findings? We see that the MMT seems to be able to provide correct predictions in terms of RTs for NVC responses for valid but does not for invalid syllogisms. The systemic hypotheses proposing an early NVC aversion and a later mental depletion seem to be able to explain why cognitive theories sometimes fail to predict NVC responses correctly: Yet, cognitive theories do not yet take such processes into account. It is noteworthy however, that the systemic hypotheses are unable to explain some of the results found in the present study. Whereas the cognitive theories do at least predict an effect of NVC-responses on the RTs, two of the systemic hypotheses do not necessarily propose higher RTs for such trials. Additionally, one of the systemic hypotheses predicted the main effect of validity on NVC responses.

In summary, with regard to the proposed theories, we see that the cognitive theories seem to be able to provide correct predictions of NVC responses for valid but sometimes not for invalid syllogisms. The strong dependencies on the validity of a syllogism as well as differences over the time-course of an experiment suggest that there are also some other cognitive processes taking place within the individual. The systemic hypotheses can account for some of these effects complementing the cognitive theories. We can conclude that there may indeed be an initial bias against an NVC response, which highly differs between individuals. Hence, more analysis are necessary to analyze the interplay between existing cognitive reasoning theories and possible systemic hypotheses to increase the correct prediction rate of when people answer NVC. Indeed, in parallel to this work, we were able to demonstrate that by attaching heuristic rules for predicting NVC to cognitive models of syllogistic reasoning, their performance can increase up to 20 % on average (Riesterer, Brand, Dames, & Ragni, in press). Last, the results also highlight that logical correctness need to be used with caution when analyzing syllogistic reasoning data due to the unproportional weight of NVC responses: Such analyses should always consider the validity of the problems.

Acknowledgements

This paper was supported by DFG grants RA 1934/3-1, RA 1934/2-1 and RA 1934/4-1 to MR.

References

Achourioti, T., Fugard, A. J., & Stenning, K. (2014). The empirical study of norms is just what we are missing. *Frontiers in Psychology*, *5*, 1159.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390 - 412. (Special Issue: Emerging Data Analysis)

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255 - 278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

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

Copeland, D. E. (2006). Theories of categorical reasoning and extended syllogisms. *Thinking & Reasoning*, *12*(4), 379–412.

Costa, A., Saldanha, E.-A. D., Hölldobler, S., & Ragni, M. (2017). A computational logic approach to human syllogistic reasoning. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (p. 883-888).

Hardman, D. K., & Payne, S. J. (1995). Problem difficulty and response format in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 945–975.

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

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54-69.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–57.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

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

Ragni, M., Riesterer, N., Khemlani, S., & Johnson-Laird, P. (2018). Individuals become more logical without feedback. In T. Rogers, M. Rau, J. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1584–1589). Austin, TX: Cognitive Science Society.

Riesterer, N., Brand, D., Dames, H., & Ragni, M. (in press). Modeling human syllogistic reasoning: The role of "no valid conclusion". In A. Goel, C. Seifert, & A. Arbor (Eds.), *Proceedings of the 41th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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

Schmeichel, B. J., & Vohs, K. D. (2018). Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. In *Self-Regulation and Self-Control* (pp. 318–347). Routledge.

Shannon, C. E., & Weaver, W. (1963). The mathematical theory of communication. 1949. *Urbana, IL: University of Illinois Press*.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Wetherick, N. E., & Gilhooly, K. J. (1995). Atmosphere, matching, and logic in syllogistic reasoning. *Current Psychology*, *14*(3), 169-178.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460.