Generalizing Syllogistic Reasoning: Extending Syllogisms to General Quantifiers

Daniel Brand[1] **(daniel.brand@hsw.tu-chemnitz.de)**

Maximilian Mittenbühler[2] **(max.mittenbuhler@gmail.com)**

Marco Ragni[1,2] **(marco.ragni@hsw.tu-chemnitz.de)**

[1] Predictive Analytics, Chemnitz University of Technology, Germany,
[2] Department of Computer Science, Albert Ludwig University of Freiburg

Abstract

Syllogistic reasoning is one of the oldest domains of reasoning research and has made great advances in understanding and modeling human reasoning processes. However, the field was mostly focused on a traditional set of quantifiers originating in first-order logic, thereby neglecting the large variety of quantifiers humans use when engaging in reasoning in their everyday life. The present work makes three main contributions: (I) we conducted a study yielding a dataset covering all traditional syllogisms and tasks containing generalized quantifiers "most" and "most not", providing a starting point for existing theories and models to transition to generalized quantifiers. (II) based on the dataset, we analyze the impact that the additional quantifiers have on the reasoning behavior. (III) We investigated the reasoning behavior with respect to the difference between traditional and generalized quantifiers, gaining insights into some of the peculiarities of the domain of generalized syllogisms.

Keywords: syllogistic reasoning; generalized quantifiers

Introduction

The first investigations of syllogistic reasoning by Störring (1908) took place a century ago, which makes syllogistic reasoning one of the oldest domains of reasoning research. A syllogism consists of two quantified statements (called *premises*) connecting three terms A, B and C via the middle term B and pose the task to derive a conclusion about the relation between the end-terms A and C. In traditional syllogisms, the quantifiers used in the premises are one of the four quantifiers underlying first-order logic: *all* (A), *some* (I), *some not* (O), or *none* (E).

The respective letters in parentheses are classical abbreviations commonly used in the field (e.g., Pfeifer, 2006), which will also be used throughout this article. Furthermore, the terms in a syllogism can be arranged in four different ways, which are referred to as *figures*. In this article, we use the definition of figures used by Khemlani & Johnson-Laird (2012), which is shown in the following table:

	Figure 1	*Figure 2*	*Figure 3*	*Figure 4*
Premise 1	A-B	B-A	A-B	B-A
Premise 2	B-C	C-B	C-B	B-C

The abbreviations of the quantifiers together with the Figure allow to abbreviate the type of syllogism. As an example, consider the following two assertions, which form the syllogism *II1*:

(1) Some cognitive theories are predictive.
(2) Some predictive theories are field-changing.

What, if anything, follows from these two premises?

Most people do infer from this syllogism that some cognitive theories are field-changing (Khemlani & Johnson-Laird, 2012). While this fits our background knowledge and the findings in experiments, it is not logically valid, which means that counter-examples can be found that contradict the conclusion. However, logic is a normative framework, but does not necessarily describe the way how humans reason about these tasks. Explaining the processes leading to this deviation from logic lies at the core of most theories of syllogistic reasoning. A variety of theories for syllogistic reason exists, which can predict the overall response of the participants on aggregated data by up to 85% (Khemlani & Johnson-Laird, 2012), which means that existing theories can —at least to some extent —explain reasoning for the syllogistic problems.

In total, the domain of syllogistic reasoning consists of a set of 64 tasks, which can be formed by combining four quantifiers in two premises with four possible arrangements. The fact that the domain is finite and is small enough to make collecting complete datasets covering all tasks feasible makes it not only a good fit for model development, but also detailed evaluations (e.g., Riesterer et al., 2020a) and in-depth analyses of patterns in the data (Brand, Riesterer, Dames, & Ragni, 2020).

However, most work focused on the traditional Aristotelian syllogisms which only consider four quantifiers, while humans consider a variety of additional quantifiers in their everyday reasoning (e.g., Barwise & Cooper, 1981; Geurts, 2003). This leads to a core question of this work: Are the findings still valid when dealing with an extended set of *generalized* quantifiers, i.e., the quantifiers such as *most*? From this question, two important points arise:

First, most experimental findings on the basis of the traditional syllogism considered a limited set of quantifiers not only for the premises (All, Some, Some not, None), but also for the possible conclusions that the participants could reason about. To this end, it is not clear if participants would have shown a different reasoning behavior if additional options were available. Even for the 64 traditional syllogisms, the findings could be heavily influenced by the selection of quantifiers.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Second, the processes of human reasoning might also change depending on the type of the quantifiers (e.g., when leaving traditional quantifiers), which would decrease the significance of theories and models explaining reasoning for first-order logic tasks as accounts of general human reasoning.

The present work aims at tackling both points by conducting a study that extends the traditional syllogism by the generalized quantifiers *most* (abbreviated by T) and *most not* (D). This allows us to obtain a complete dataset, that consists of the full set of syllogisms that are possible with the new quantifiers, including the 64 traditional syllogism. This allows us to investigate the differences between the traditional and generalized quantifiers, assessing the question if the domains of the traditional and generalized quantifiers can be considered to be independent. The resulting dataset can also serve as a starting point for evaluating or extending models and theories on an extended set of quantifiers. Furthermore, we will tackle the question, if the reasoning processes for both domains are different and provide a first overview of the effects occurring when transitioning from first-order logic to generalized quantifiers.

In the following, we will first introduce relevant work covering the traditional syllogisms as well as generalized quantifiers. Second, our study and the resulting dataset is presented. Third, we describe our analyses and results. Finally, we will discuss our findings and implications for the field of syllogistic reasoning research.

Related work

Over the years, models and theories for syllogistic reasoning cover the wide range from heuristic, probabilistic, logical and model-based approaches (Khemlani & Johnson-Laird, 2012). While the number of theories is still increasing (Tessler et al., 2022; Brand, Riesterer, & Ragni, 2020), the recent years showed a shift of focus towards model evaluation and benchmarking (Riesterer et al., 2018, 2020b), allowing to compare models on fair grounds. However, evaluations fall and rise with the quality of the underlying data. While comprehensive datasets for traditional syllogisms are available, datasets that include generalized quantifiers are rare, hindering the transition of modeling and model evaluations to generalized quantifiers.

Pfeifer (2006), Evans (2002), and others have criticized the limitations of traditional quantifiers with respect to everyday communication and reasoning. The universal quantifiers, i.e., *None* (E) and *All* (A) are too strict as they do not allow for any exceptions. In contrast, *Some* (I) and *Some ... not* (O) are considered too weak, as they are already satisfied by a single element in the set (Pfeifer, 2006). In contrast, everyday human reasoning is "based [. . .] on beliefs, in which there are varying degrees of confidence" (Evans, 2002, p. 980). Given its relevance there have been only few experimental investigations of generalized quantifiers such as most or few (Oaksford & Chater, 2001; Chater & Oaksford, 1999; Pfeifer, 2006) and few extensions and evaluations of cognitive theories (Ragni et al., 2014). While we refer for the traditional syllogisms to the article by Khemlani & Johnson-Laird (2012), we will briefly introduce the generalized quantifiers *most*, which has been investigated in about all papers for generalized quantifiers (Geurts, 2003).

How is the quantifier *most* interpreted formally? A logical interpretation of the quantifier $Most(A,B)$ for finite sets A and B is true, if the elements that are in both A and B at the same time is greater than the number of elements that are only in A, but not in B. Formally, it holds $|A \cap B| > |A - B|$, with $|\cdot|$ being the size or the number of their elements (Westerståhl, 1989; Novák, 2008), but see for other interpretations (e.g., Hackl, 2009). It is important to notice that the quantification over the size of sets is formally not possible in first order logic (and requires at least a fragment of second order logic).

How is the quantifier *most* interpreted by humans? A corollary from the above that "Most S are P is to say that there are more S that are P than S that are not P" has been used in experimental investigations (Pfeifer, 2006). Thompson (1982) argues that "Few S are not P" makes a strong enough claim that it would be invalid to infer it from the weaker claim "Most S are P". The definition used by Chater & Oaksford (1999) suggests that the meaning should be given in terms of constraints on the conditional probability, leading to "Most S are P" being interpreted as the probability of P given S being high (but less than 1, essentially excluding *all* as a possible meaning). This is also in line with the Gricean implicature that states that we are as informative as required but not more informative (Geurts, 2010, p.11), therefore excluding *all* from *most* in language. Furthermore, according to Chater & Oaksford (1999), *few* and *most* are used as inverse relations. However, Ragni et al. (2017) demonstrated that *few* might not be perceived that way by humans. Finally, Newstead et al. (1987) investigated the interpretation of quantifiers in rating scales, finding a dependency of the interpretation on the set size. This could also be problematic for syllogistic tasks, as the set sizes are usually not defined, potentially leading to variance due to different assumptions about the set sizes.

Study

For the 64 first-order logics based syllogisms (referred to as *traditional syllogisms* in this article), datasets containing the response behavior for all syllogisms and all possible conclusions exist (e.g., Dames et al., 2020). These datasets allow for a rich analysis of the response patterns (e.g., Brand, Riesterer, Dames, & Ragni, 2020), that would not be possible without a complete dataset. In this work, we aim at extending the traditional domain to generalized quantifiers in a way that allows existing models and theories for the traditional syllogisms a smooth transition. Therefore, it is important to also provide a comprehensive and complete dataset that also contains the traditional syllogism. The inclusion allows us to investigate the impact that additional options for conclusions have on the reasoning process and it provides a starting point for extending existing syllogistic theories and models.

Experimental setup and dataset

However, while the traditional syllogisms consist of a well-defined, finite set of 64 tasks, syllogisms containing generalized quantifiers will always be an arbitrary selection of tasks and/or quantifiers. Even when restricting the quantifiers to commonly used ones (e.g., most, few, many), the number of possible tasks grows exponentially with each additional quantifier. In order to obtain a complete dataset that is comparable to the datasets available, we only considered the generalized quantifier *most* and the negated form, *most not*. By choosing *most not* over *few*, we aimed at avoiding additional issues with the exact interpretation, as it is still debated if they offer a comparable meaning. Additionally, it brings the advantage of being more consistent to the handling for *some* and *some not*. As mentioned above, the traditional quantifiers from the 64 syllogisms featuring the first-order logic quantifiers to generalized quantifiers were also included, leading to a total of 6 quantifiers and 144 syllogisms (64 of which were traditional syllogisms).

With the inclusion of generalized quantifiers we leave the traditional frame of first-order logic and additional effects might come into play. For example, quantifiers have been shown to slightly vary in their interpretation based on the set sizes (Newstead et al., 1987). Since the set sizes are not defined in syllogistic reasoning, the interpretation of quantifiers that do not have a classical logical interpretation might introduce an uncontrollable source of variance, as participants could base their reasoning on an arbitrary set size. This brings up another reason for choosing *most*: Its interpretation refers to the majority of a set (e.g., more than half) and is therefore presumably more invariant to assumptions about the set size. Additionally, it allows to determine the logical correctness and the possibility to reason about it while still being distinct to *some* and *all* (e.g., syllogisms like TT4 are valid, while II4 is not and others like TT1 are invalid, while AA1 is valid). Other common options like *many* on the other hand are ambiguous as they can be understood as *many of*, resulting in a similar meaning to *most*, but can also refer to the total number of elements (Thompson, 1982).

We acquired data from 31 participants (female: 16, mean age: 36.8, SD age: 14.0) using the platform Prolific[1]. Each participant responded to all 144 syllogisms over 3 sessions, resulting in a total of 4464 data points. Participants were presented with one syllogism at a time and were asked to select the conclusion from the set of all 13 possible conclusions (6 quantifiers with 2 possible ways to order the end terms, yielding 12 possible conclusions, as well as the possibility that there is no valid conclusions (NVC)). The order in which the syllogisms were shown was randomized. All syllogism contained professions as subjects in order to avoid biases. The professions differed in each syllogism and were randomized for every participant. To reserve participants' attention, the experiment was split into three separate sessions, which they could complete at any time. They took about 45 minutes on average to complete each session. In addition to their response to the syllogisms, we obtained demographic, reaction time, and situational data from participants (e.g., what strategy they had used to respond, how attentive they had been during the experiment, to what extent they had guessed their responses). Additionally, they were asked about their interpretation of the quantifier *most*, i.e., if it also includes *all*.

[1]https://www.prolific.co/

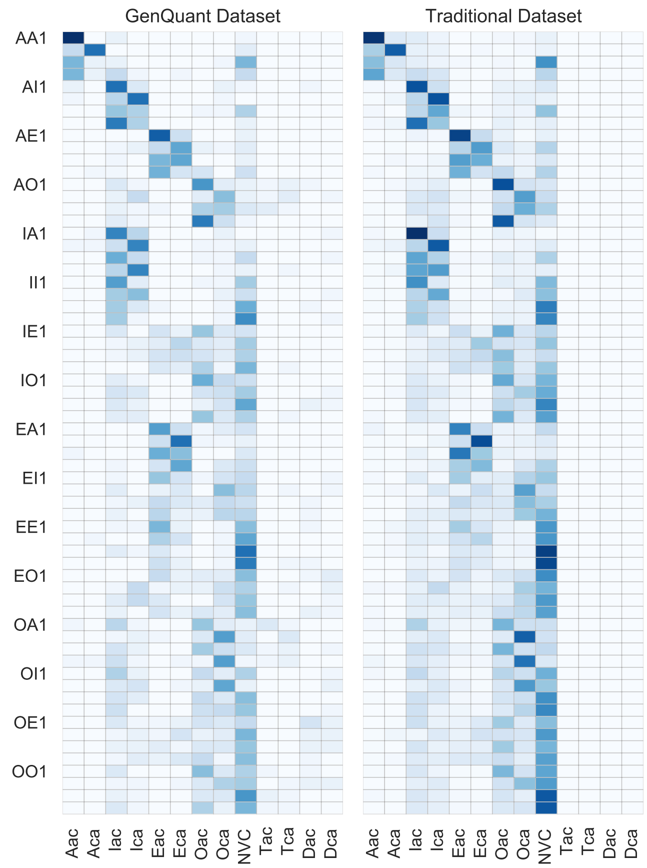

Figure 1: Comparison of the patterns on the traditional 64 syllogisms between the general quantifiers dataset (left) and the Ragni-2016 dataset, which did only contain the traditional syllogisms and response options (right).

Analysis

Using participants' responses to the full set of syllogisms based on our six quantifiers, we firstly investigated if traditional syllogisms can be examined as an independent domain from generalized syllogisms. To this end, we assessed how participants' response patterns were affected by the introduction of generalized quantifiers. The dataset and anylsis is available on GitHub[2]. Secondly, we explored the effects of adding generalized quantifiers by analyzing their order of

[2]https://github.com/Shadownox/cogsci-2022-genquant

informativeness and by comparing the difficulty of the syllogisms. We will describe each of these analyses in turn.

Are traditional syllogisms an independent domain?

The additional response options given by generalized quantifiers could affect the responses to traditional syllogisms. To investigate this, we compared the response patterns of our dataset to the response patterns found in the Ragni-2016 dataset that can be obtained from the Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) Framework[3]. The dataset contains the responses of 139 participants to all 64 traditional syllogisms (with correspondingly fewer response options). Figure 1 shows the average response patterns for the traditional syllogisms in the present dataset compared to the Ragni-2016 dataset. As the Ragni-2016 dataset only had the 9 traditional response options available, the columns for the generalized quantifiers are empty by definition. Generally, the difference between both datasets is low, with both datasets clearly showing the same general patterns. In order to quantify the difference, we calculated the root mean squared error (RMSE) between the response distributions for each syllogism. The mean RMSE across all tasks was 0.04, indicating that the results on both datasets can be considered to be comparable. It becomes apparent that the additional response options do not seem to have an impact on the general response behavior. While there are some responses using the additional quantifiers, they are not substantial and could be attributed to guessing behavior. In the following analysis, we aim at providing detailed insights at the interactions between quantifiers in the premises and responses, which also allows to put these responses into perspective.

Next, we examined how the quantifiers in the premises and in the responses interact when generalized quantifiers are added. Figure 2 depicts a crude view of participants' answering patterns for each group of syllogisms (i.e., ignoring figures and the order of premises). We further divided syllogisms and responses into universal, existential, or generalized, depending on the quantifiers. As is evident, participants mostly responded with the same type of quantifier that was already present in the premises (83.6% of responses). This is specifically true for syllogisms containing only universal quantifiers (86.7%), existential quantifiers (91.7%), or a mixture thereof (91.7%). This corroborates the notion that adding generalized quantifiers does not influence the answers given to traditional syllogisms. Only when generalized quantifiers were involved, participant noticeably swayed from this pattern, mostly by responding with existential quantifiers even if these were not present (29.6% for a mixture of universal and generalized; 37.6% for generalized only). In sum, these two findings indicate that the traditional syllogisms, albeit being a subset of the syllogism with generalized quantifiers, can be investigated independently. This is especially important, as most findings in the field of syllogistic reasoning research

[3] https://github.com/CognitiveComputationLab/ccobra

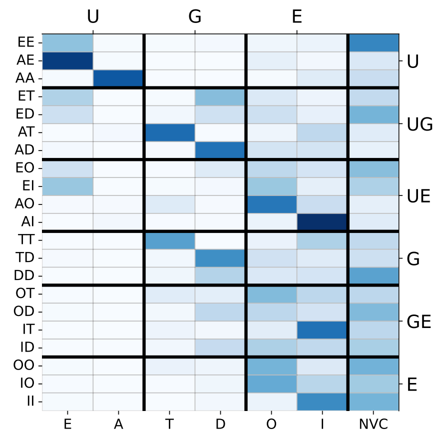

Figure 2: Participants' responses for all syllogisms collapsed across figures and orders of premises. A darker color indicates a higher number of responses. Syllogism groups are shown on the right, responses on the bottom. Quantifier types (universal: U, existential: E, generalized: G) are shown on the right and top for premises and responses, respectively.

were based on the traditional syllogisms only and would have been severely affected if additional quantifiers had altered the results significantly.

Effects on correctness

Although traditional syllogisms seem to be largely independent from the syllogisms with generalized quantifiers, the processes when solving tasks from both domains can still be similar. In the following, we aim at providing additional insight into the reasoning behavior for both domains, by focusing on the correctness of the given responses. For the assessment of the correctness in the analyses in this work, we settled with the interpretation that *most A are B* refers to the majority, i.e., more than half, of the elements in A are B. Furthermore, it does include the case that *all* elements of A are B. Table 1 shows the correct conclusions for all valid syllogisms that differ to the traditional syllogism (for invalids, the correct answer is NVC; for the remaining syllogisms see Khemlani & Johnson-Laird, 2012).

Figure 3 shows the mean correctness for the different tasks broken down by the type of the quantifiers in the premises. It becomes apparent that the performance substantially drops for syllogisms that include a generalized quantifier. In the following, we assess three possible explanations for this.

First, reasoning with the quantifier *most* might be less common, while traditional syllogisms and first-order logic are more common. This could imply that participants were more familiar with the traditional tasks. However, participants were asked if they participated in similar experiments, which only 1 out of the 31 participants answered affirmatively.

Second, the results might be an artefact of the tasks. Invalid syllogisms, which only have a single solution (namely that there is no valid conclusion possible), are often considered to

Table 1: Valid syllogisms and the respective conclusions. The direction of the conclusion is indicated by *ac* and *ca*. Only syllogisms that are not part of the traditional set have differing conclusions are included.

Task	Conclusions	Task	Conclusions	Task	Conclusions
AA1	Aac, Iac, Ica, Tac	EA2	Eac, Eca, Oac, Oca, Dac, Dca	TE1	Oac, Dac
AA2	Aca, Iac, Ica, Tca	EA3	Eac, Eca, Oac, Oca, Dac, Dca	TE2	Oac
AE1	Eac, Eca, Oac, Oca, Dac, Dca	ET1	Oca	TE3	Oac, Dac
AE3	Eac, Eca, Oac, Oca, Dac, Dca	ET2	Oca, Dca	TE4	Oac
AT2	Iac, Ica, Tca	ET3	Oca, Dca	TT4	Iac, Ica
AT4	Iac, Ica	ET4	Oca	TD4	Oac
AD3	Oca, Dca	TA1	Iac, Ica, Tac	DA3	Oac, Dac
AD4	Oac	TA4	Iac, Ica	DA4	Oca
				DT4	Oca

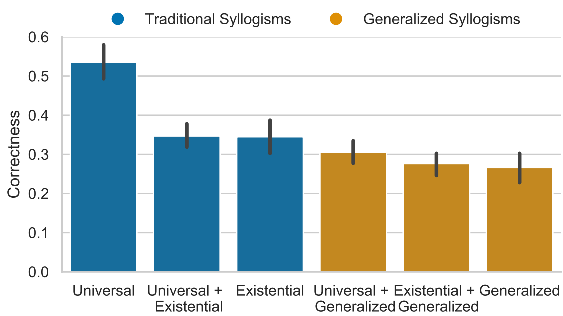

Figure 3: Mean correctness of participants' responses for different syllogisms broken down by quantifier types of the premises (Universal, Existential, Generalized).

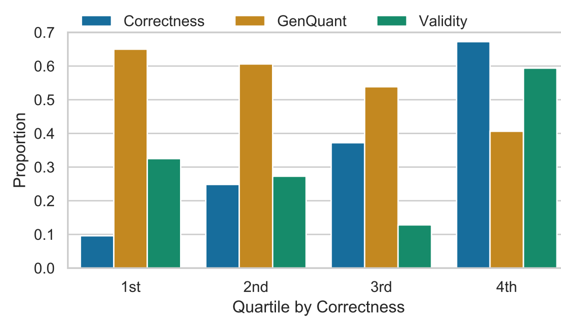

Figure 4: Proportion of correctly solved tasks, occurrence of generalized quantifiers and valid syllogisms for different quartiles based on the correctness.

be more difficult due to a variety of reasons ranging from the necessity to perform a exhaustive search for counterexamples to general biases against the response option (e.g., Ragni et al., 2019; Brand, Riesterer, & Ragni, 2020). However, the addition of the quantifiers *most* and *most not* introduced a total of 80 new tasks, out of which 61 are invalid (while there are 37 invalid syllogisms for the 64 traditional syllogisms). Given the potentially more challenging nature of invalid tasks, this could explain the difference. To investigate this, we divided the tasks into quartiles based on the mean performance for the respective task and assessed the proportion of generalized quantifiers and valid tasks in the quartiles (see Figure 4). While the easiest tasks (4th quartile) indeed seem to confirm the assumption that the correctness largely depends on the validity, the remaining groups do not. Instead, mostly the relation between the proportion of generalized quantifiers and the correctness becomes apparent, while the validity seems to be of secondary importance.

Third, participants might use a different interpretation for *most*, which in turn leads to a systematic misjudgement of their performance. When asked if the meaning of *most* also includes *all*, 27 participants responded that it does not, while only 2 participants agree that most can also include all (2 participants responded that they do not know). This can explain why certain tasks containing generalized quantifiers have very low correctness values (e.g., AD2, which was not solved correctly by any participant). These tasks seem to be valid, as the counterexamples for the conclusions require to consider a universal meaning of the quantifier. To assess this explanation, we also considered the correct conclusions under the assumption of the alternative interpretation. To achieve this, we replaced the quantifier, essentially mapping the task to another task. For example, for the task *AT2* with the quantifier *most*, the alternative interpretation implies that some elements are excluded, which means that we also considered the result for the quantifier *some not*, respectively the task *AO2*. The same principle holds for existential quantifiers, where *some* could also imply *some not*. Note that this means that we could also continue the replacement from *most* over *some not* to *some*, but as it is already implied by *most*, it wouldn't yield any additional possible conclusions. Additionally, it is important to note that in the case of generalized quantifiers and existential quantifiers occurring together, the replacement is not necessary, as the task is already invalid and there is no possibility leading to a valid task. Figure 5 shows the proportion of correctly and incorrectly solved tasks, as well as tasks that were solved correctly by relying on the misinterpreted quantifiers, as determined by the described procedure. It be-

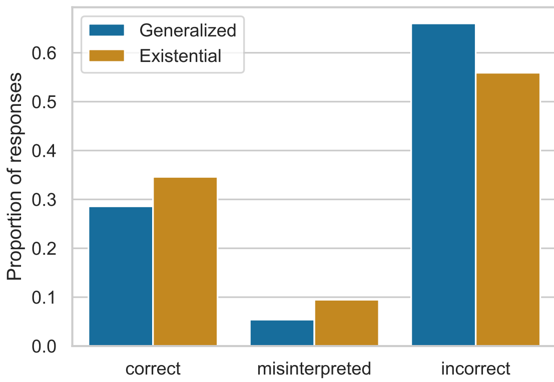

Figure 5: Proportion of tasks solved correctly with and without considering the misinterpreted meaning of quantifiers, and incorrectly solved tasks broken down by the occurrence of generalized and existential quantifiers. Note that *existential* only considers the traditional syllogism.

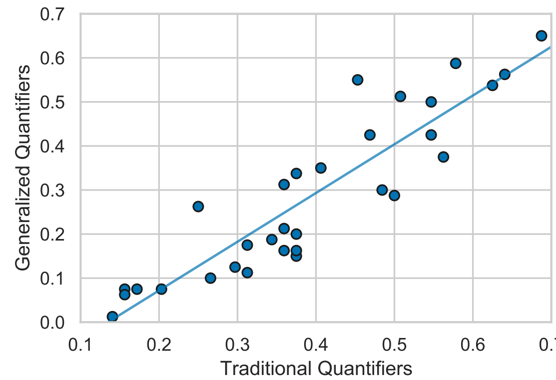

Figure 6: Participants' correctness on traditional quantifiers against generalized quantifiers. Each point represents an individual participant.

comes apparent, that in most cases, errors were not caused by the alternative interpretation of the quantifiers. While this does not mean that misinterpretations did not play a role in the reasoning processes, they can be ruled out as the primary cause of errors.

The difference in correctness can be interpreted in two ways: On the one hand, the tasks might simply be more demanding, increasing the likelihood of mistakes. On the other hand, they might be processed differently. In order to shed some light on these options, we investigated the differences between the performance on tasks with generalized and traditional quantifiers on an individual level: If generalized quantifiers are processed differently, participants that solved them correctly won't necessarily solve the traditional syllogisms and vice versa. Figure 6 shows the mean correctness for each participant on traditional syllogisms plotted against the correctness on syllogisms featuring the generalized quantifiers. The correctness for both types of syllogisms seems to be tightly coupled, which also shows in a high correlation between both (Spearman $r = 0.91$, $p < 0.001$). This indicates that the ability to solve the traditional tasks seems to extend to the generalized quantifiers and therefore supports the assumption that the same processes are used for both domains.

Discussion

With this work, we aimed at providing a starting point for advancing the understanding of syllogistic reasoning to the domain of generalized quantifiers. To our knowledge, no comprehensive and complete dataset exists that contains generalized quantifiers as well as all traditional syllogisms. While some theories exist that are already capable of dealing with those (e.g., PHM Chater & Oaksford, 1999; Oaksford & Chater, 2001) the lack of data prevented a rigorous evaluation of the theories and hindered in-depth analyses of the domain. We conducted a study which extended the traditional syllogisms by the quantifiers *most* and *most not*, obtaining a first complete dataset that allows to assess the transition between the traditional and generalized quantifiers. Our analysis showed that the addition of conclusions with generalized quantifiers did not have an impact on the reasoning behavior for the 64 traditional tasks. This finding is important for two reasons: First, if the reasoning behavior had been affected, it would have been necessary to reconsider findings based on the traditional tasks alone, as they could have been an artefact of the task design. Second, the independent nature both domains allows future studies to exclude the traditional syllogisms when investigating generalized quantifiers, which is greatly beneficial in a domain that has already a large number of tasks (which quickly becomes experimentally unfeasible due to the exponential growth when adding new quantifiers).

Furthermore, we investigated general properties of the two domains, especially with respect to the reasoning performance. We found that generalized tasks seemed to be substantially more challenging. Our analyses could rule out that the lower performance was due to misinterpretations or the high proportion of invalid syllogism in the generalized tasks. This leads to the question, if different processes are used when solving generalized tasks, or if the tasks are just more difficult due to their innate properties. Although we found that the ability to correctly solve the tasks is highly correlated between the traditional and the generalized tasks, it is ultimately a question that needs to be targeted by cognitive models and theories. We hope that our work can serve as a basis for further research extending the domain of syllogistic reasoning to a wider range of quantifiers that better reflect everyday reasoning and provide an insight into different facets of logical reasoning.

Acknowledgements

Support to MR by the German Research Foundation, DFG (RA 1934/4-1, RA1934/8-1) is gratefully acknowledged.

References

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence* (pp. 241–301). Springer.

Brand, D., Riesterer, N., Dames, H., & Ragni, M. (2020). Analyzing the differences in human reasoning via joint nonnegative matrix factorization. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3254–3260). Cognitive Science Society.

Brand, D., Riesterer, N., & Ragni, M. (2020). Extending TransSet: An individualized model for human syllogistic reasoning. In T. C. Stewart (Ed.), *Proceedings of the 18th International Conference on Cognitive Modeling* (pp. 17–22). University Park, PA: Applied Cognitive Science Lab, Penn State.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258.

Dames, H., Schiebel, C., & Ragni, M. (2020). The role of feedback and post-error adaptations in reasoning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3275–3281). Cognitive Science Society.

Evans, J. S. B. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*(3), 223–251.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural language semantics*, *17*(1), 63–98.

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.

Newstead, S., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, *18*(3), 178-182. doi: https://doi.org/10.1016/0003-6870(87)90001-9

Novák, V. (2008). A formal theory of intermediate quantifiers. *Fuzzy Sets Syst.*, *159*, 1229-1246.

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

Pfeifer, N. (2006). Contemporary syllogistics: Comparative and quantitative syllogisms. In G. Kreuzbauer & G. J. W. Dorn (Eds.), *Argumentation in Theorie und Praxis: Philosophie und Didaktik des Argumentierens* (pp. 57–71). Wien: LIT.

Ragni, M., Dames, H., Brand, D., & Riesterer, N. (2019). When does a reasoner respond: Nothing follows? In A. K. Goel, C. M. Seifert, & C. Freska (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2640–2646). Montreal, QB: Cognitive Science Society.

Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., & Tse, A. P. P. (2017). Formal nonmonotonic theories and properties of human defeasible reasoning. *Minds I& Machines*, *27*, 79–117.

Ragni, M., Singmann, H., & Steinlein, E.-M. (2014). Theory comparison for generalized quantifiers. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1228–1233). Cognitive Science Society.

Riesterer, N., Brand, D., & Ragni, M. (2018). The predictive power of heuristic portfolios in human syllogistic reasoning. In F. Trollmann & A.-Y. Turhan (Eds.), *Proceedings of the 41st German Conference on AI* (pp. 415–421). Berlin, Germany: Springer.

Riesterer, N., Brand, D., & Ragni, M. (2020a). Do models capture individuals? Evaluating parameterized models for syllogistic reasoning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3377–3383). Toronto, ON: Cognitive Science Society.

Riesterer, N., Brand, D., & Ragni, M. (2020b). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, *12*(3), 960–974.

Störring, G. (1908). *Experimentelle Untersuchungen über einfache Schlussprozesse*. W. Engelmann.

Tessler, M. H., Tenenbaum, J. B., & Goodman, N. D. (2022). Logic, probability, and pragmatics in syllogistic reasoning. *Topics in Cognitive Science*, *n/a*(n/a). doi: https://doi.org/10.1111/tops.12593

Thompson, B. (1982). Syllogisms using "few", "many", and "most". *Notre Dame Journal of Formal Logic*, *23*, 75–84. doi: doi:10.1305/ndjfl/1093883568

Westerståhl, D. (1989). Aristotelian syllogisms and generalized quantifiers. *Studia Logica*, *48*(4), 577–585. doi: 10.1007/BF00370209