

How people do relational reasoning? Role of problem complexity and domain familiarity

Shikhar Kumar

School of Information: Science, Technology and Arts, University of Arizona, Tucson, AZ,

Iliano Cervesato

School of Computer Science, Carnegie Mellon University, Doha, Qatar

Cleotilde Gonzalez

Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA

Abstract

The goal of this paper is to study how people do relational reasoning, such as selecting the grade of all students in a class with GPA (Grade Point Average) greater than 3.5. Literature in the field of psychology of human reasoning offer little insight as to how people solve relational problems. We present two studies that look at human performance in relational problems that use basic relational operators. Our results present the first evidence towards the role of problem complexity on performance as determined by the accuracy and discrimination rates. We also look at the role of familiarity with tabular representation of information, as found in spreadsheets for example, and other factors for relational reasoning, and show that familiarity does not play a significant role in determining performance in relational problem solving, which we found counterintuitive.

Keywords: Relational reasoning, Problem solving, Spreadsheet, Domain familiarity, Problem complexity

1. Introduction

Nowadays, data are more easily accessible than ever, yet support for deriving interesting consequences from base data is often unavailable, too expensive, or too technical for many users. For example, a student may have access to prerequisite listings and expected offering dates of courses but have no way to sieve through possible course sequences unless the college provides a dedicated tool. Similarly, an investor may know the instruments held in his mutual fund portfolio but have no easy way to unravel them and reveal his exposure to a specific industry or company. In all cases, manually inferring useful information from raw data is time consuming and error prone, a situation that often results in bad decisions, suboptimal plans, or missed opportunities. In fact, there is currently no simple and general application that empowers users to compute useful inferences on raw data.

Cervesato (2007; 2013) addressed this problem by drawing inspiration from a type of automated data inference that is immensely popular: the spreadsheet. Applications such as Microsoft Excel and others are readily available and allow users to routinely perform complex custom calculations on numerical data. The spreadsheet's clever interface makes it easy to use

productively with little or no training. However, none of the above data manipulation problems is expressible in today's spreadsheets. The approach investigated in the cited work, which was dubbed NEXCEL in (Cervesato, 2007), remedies this situation by extending the spreadsheet paradigm to enable users to define useful forms of inference among their data. It allows the student, for example, to download data about course prerequisites and offerings into his favorite spreadsheet, and write a "formula" that calculates all possible course sequences. The investor can similarly see the individual stocks in his portfolio and determine his actual exposure.

These "formulas" combine not numbers but *relations* (for example the relation that associates courses to each of their prerequisites, or the relation between mutual funds and publicly traded companies). Just like traditional spreadsheets leverage the ability of their users to capture numerical inferences using numerical formulas, NEXCEL asks users to express *relational reasoning* using these relational formulas. However, little is known about how people do relational reasoning. Clearly some people are very good at it (e.g., database programmers). But how natural is it for the rest of us? Which relational constructs do humans find easy to use? Which ones lead us to make more mistakes? How should a language of relational formulas be constructed to capture most immediately the way we do relational reasoning?

We did not find answers to these questions in the literature. The closest studies we are aware of are Johnson-Laird's analysis of how people fare with various forms of logical inferences (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991; Johnson-Laird & Byrne, 1992; Johnson-Laird, 2006) and Oaksford and Chater's probabilistic approach to human reasoning (Oaksford&Chater, 2001; Oaksford&Chater, 2007; Oaksford&Chater, 2009).

The analysis of human performance in reasoning tasks have shown that people make large and systematic errors, which are not random (Evans, 1993; Manktelow, 1999), suggesting that humans might be irrational (Stich 1985; Stein 1996). These observations have led to numerous studies, with several formal models of human reasoning. Some of the most well known approaches involve the comparison of human performance against formal logic.

In the logical analysis of human reasoning, two major approaches have been utilized: mental-logic approach (Rips, 1994) and mental-model approach (Johnson-Laird, 1991). Both these approaches argue that the systematic deviations from logic in deductive inference tasks represent unavoidable performance errors, which stem from limited working memory. Given humans have limited working memory and other cognitive abilities, it restricts their reasoning abilities. Thus, in principle humans are rational but in practice they are constrained by cognitive limitations.

In contrast to logic based approaches, a more recent approach to the analysis of human reasoning is the probabilistic approach (Oaksford&Chater, 2007). It posits that everyday reasoning is probabilistic and the reason why people make errors in logical tasks conducted in the laboratories is because they generalize these everyday strategies to the laboratory. Oaksford and Chater argue that logic is inadequate to account for everyday reasoning and probabilistic approach is more promising (Oaksford&Chater, 2001). This approach has been applied to several core areas of the psychology of reasoning: conditional inference (Oaksford et al., 2000; Schroyens, 2000a; Schroyens, 2000b), syllogistic reasoning (Chan & Chua, 1994; George, 1997; Liu et al., 1996; Stevenson & Over, 1995; George, 1999), and Wason's selection task (Oaksford&Chater, 1994; Oaksford&Chater, 1998). In all these areas it has been shown that

probabilistic approach offers a better explanation to human performance than more traditional normative approaches.

Here we will not argue which approach is better; rather we will consider what both the logical and probabilistic approaches to human reasoning have to offer with regards to our understanding of relational inference. Both offer two different explanations of the performance of humans in inference tasks: 1) Limitation of cognitive abilities 2) People employ strategies from everyday reasoning in the laboratory. We posit that limitations of cognitive abilities will play a significant role in relational inference. However, we are not quite sure if the strategies from everyday reasoning will be utilized in relational inference and how they might affect the performance.

Answering these questions is critical to the development of tools like NEXCEL, and more generally to understand the ways in which humans carry out relational inferences, how relational inferences compare to numerical inferences; and how they would perform both types of inferences with spreadsheet capabilities. To this end, we have designed a series of studies whose purpose is to answer precisely these questions. In the present paper, we report on two studies that explore our subjects' ability to carry out the most basic forms of relational inference. In the first experiment, we study how humans perform four elementary relational operations: projection, union, difference and join (which include more general forms of selection). We did so using a traditional spreadsheet as a visual proxy. The second experiment aims at investigating how more complex relational operations are resolved. We combined various simple operations to test human performance. Future experiments will explore more complex operators (e.g., recursion) and combinations (e.g., nested negations), and gauge the subjects' ability to express the relational reasoning patterns needed to solve a problem in a variety of relational languages.

A *relation* can be visualized as a table consisting of rows and columns. Each column, or *attribute*, holds data with a consistent meaning (e.g., the grade of a student, or the name of a mutual fund). Each row, or *record*, contains specific data in the relation, for example the name, grade, and major of a specific student in a class. Relational inference computes new relations on the basis of relations we already know, for example the students with a GPA (Grade Point Average) greater than 3.5 together with their major. Relations do not contain duplicate records. Any relational inference can be obtained by combining a small number of elementary relational operations (in the same way as any arithmetic expression is based on addition, subtraction, etc). In our experiments, we relied on four of these elementary operations: projection, union, join and difference.¹ We will now give details of these operations.

- *Projection* simply deletes some attributes from a relation (and removes any duplicate record that may ensue). For example, a professor may need to make a list of student names and their respective grades for some exam. However, he only has a full grade

¹ Traditionally, join is itself decomposed into selection and Cartesian product. The latter is rarely used in isolation, and therefore would have led to artificial experiment tasks. Selection, which is commonly used in practice, becomes a special case of join. Modern presentations include recursion as an additional relational operation. We believe it is significantly more complex than the other operations, and therefore decided to dedicate a separate study to it. Every relational inference can be expressed as a combination of union, projection, selection, Cartesian product, difference and recursion (Cervesato, 2013).

sheet of the students, with their majors, and other information. Removing the unwanted columns is a use of projection.

- *Union* combines two (or more) relations with the exact same attributes into a single relation. For example, the professor may have two grade sheets, one for each section of the same class, and may need to look at the grades of all the students in the class. This task of combining both grade sheets into one is a form of union.
- *Join* is more complex: Given two sets of records with a common attribute, join combines the records that share the same value for this attribute. For example, if a professor has a list of students and the classes they take and another list of students and the sports they play, she may need the list of all students with their respective classes and sports. Here the task can be accomplished by joining the two sets of records based on student names.
- *Difference* retains the records that are in one relation but not in a second one. Like union, both relations should have the exact same attributes. For example, a professor with separate grade sheets for the two sections of her class may want to examine the performance of the students coming to the morning section only (knowing that some students attend both the morning and the afternoon section). The operation she would use to do so is difference: she wants the record of the students in the morning section that do not occur in the afternoon section.

We expected that humans would find these relational operations easy to accomplish, and that some of the more complex operations such as join would result in more mistakes and lower accuracy. We also expected that the participant's familiarity with tabular representation of information (e.g., in spreadsheets) and other relevant topics like databases, programming, logic, mathematics, etc, would help in solving problems with more complex operations.

2. Study 1

In study 1 we focus on human performance on basic relational operations: projection, union, difference and join. We investigate the role of problem complexity and participant's familiarity on their accuracy of solving each of these types of problems.

2.1. Participants

Participants were recruited from Amazon Mechanical Turk (mTurk) for a "Relational Problem Solving" study and were compensated at a flat rate of \$0.50 for participation. No bonus was awarded for performance. A total of 398 participants completed the study. The mean age of the participants was 33.57 ($SD = 12.10$), and 57.25% of the participants were female. On average, participants took 6 minutes and 52 seconds ($SD = 3$ minutes and 16 seconds) to complete the study.

2.2. Design and Procedure

Participants solved one relational problem, which was based on one of the four basic level relational operations. The design was a between-subjects experiment where participants were assigned to one of four conditions (problems) randomly.

Each relational problem was designed in the form of a scenario: a scenario included a description of the problem, a set of tables with some sample data, and a list of 8 steps or actions

in random order that could be used to solve the problem described. The goal of the participant was to select 4 out of 8 steps and organize those steps in the correct logical order to solve the operation with a set of tables provided. The steps represented the logic of solving the problem in form of a sequence of actions to be performed on the given set of tables. Only 4 out of the 8 steps were correct, and participants were informed of this. For each participant, all eight steps were randomized such that correct and incorrect steps intermixed randomly. Participants were asked to mark the incorrect steps as “0” and correct steps with a valid number (1 – 4) representing the sequence in which the actions were to be performed for correctly solving the problem. This tested the ability of the participants to identify and order the correct steps (accuracy) and also their ability to recognize the incorrect steps (discrimination).

After assigning an order number or "0" to each of the 8 steps, participants were asked to report the difficulty level of the problem and their confidence in the solution they provided for the problem in a 7-point Likert scale. Two questions were asked: a) How difficult was the problem? b) How confident are you about your solution? Participants responded to both the questions by selecting one from 1 being “Not Difficult/Not Confident” to 7 being “Very Difficult/Very Confident”.

Then, participants were given a questionnaire with 10 questions to gauge their familiarity with spreadsheets, computer programming, mathematics, logical reasoning and problem solving. Again, we used a 7-point Likert scale, with 1 being no familiarity at all and 7 being high familiarity with spreadsheets, logical reasoning and problem solving. Details of the questionnaire are provided in the appendix.

2.3. Dependent measures

We used *accuracy* and *discrimination* as the two main performance measures. The accuracy measure evaluated the correctness of choosing and ordering the steps for a given scenario. This was computed in two parts: 1) Fraction of total number of correct steps chosen, irrespective of the order. 2) Fraction of total number of correct steps ordered in the correct order. The accuracy was computed as the average of these two parts. Hence, the accuracy for choosing and ordering the correct steps was defined as $(N_{\text{choice}} + N_{\text{order}})/2N$, where N is the total number of correct steps ($N = 4$), N_{choice} is the total number of correct steps chosen and N_{order} is the total number of correct steps correctly ordered. For example, if the participant chooses all the correct steps but does not order any of them correctly then $N_{\text{choice}} = 4$ and $N_{\text{order}} = 0$, and the accuracy rate for this participant would be 1/2 or 50%. Given there are 4 steps, we can think of each step counting 25% of the total accuracy (100%): getting the step and the order correct, both would contribute 25% of the total accuracy, while getting only the operation correct but not in the right order would contribute only 1/2 of the accuracy i.e., 12.5%.

The *discrimination* measure evaluated the participant's ability to recognize the incorrect steps for the given scenario. This was computed as the fraction of the total number of incorrect steps correctly rejected by assigning “0” in the response box. Hence, the accuracy for correctly rejecting the incorrect steps was defined as M_{reject}/M , where M_{reject} is the total number of incorrect steps correctly rejected and M is the total number of incorrect steps ($M = 4$). For example, if the participant correctly rejects two incorrect steps (assigns “0”) and falsely assigns numbers > 0 to the other two incorrect steps then $M_{\text{reject}} = 2$, and the discrimination rate for this participant would be 1/2 or 50%. Just like the accuracy rate, discrimination rate also ranged from 0 – 100%.

The familiarity score was the sum of all the responses to the experience questionnaire. Given there were a total of 10 questions with a minimum score of 1 and a maximum score of 7 for each question, the familiarity score was in the range [10, 70]. Given participants were required to provide a response to each question, the minimum score that could be achieved was 10. A score of 10 indicated lowest familiarity and a score of 70 highest familiarity. To be able to evaluate different types of experience we also partitioned the overall score into two scores: Familiarity with Excel and familiarity with other relevant areas like programming, mathematics, logic, etc. Please refer to the appendix for a list of questions belonging to each category.

2.4. Results

2.4.1. Accuracy in the different types of problems

Figure 1 represents the mean accuracy and discrimination rates for each of the four relational problems.

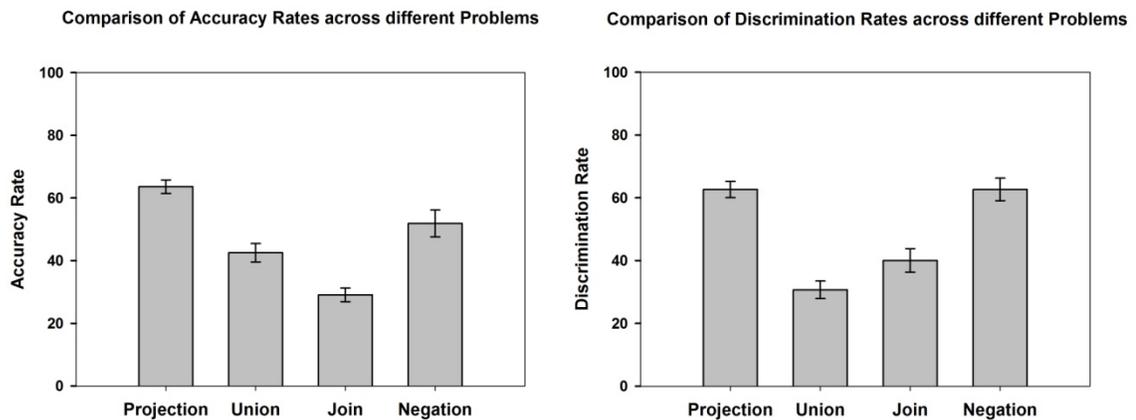


Figure 1: Left panel: Mean accuracy rates for the four different relational problems. Right panel: Mean discrimination rates for the four different relational problems. *Note:* Error bars indicate standard error of the mean.

For studying the effect of problem type on accuracy and discrimination a one-way ANOVA (Analysis of Variance) was performed with problem type (projection vs. union vs. join vs. difference) as the between-subjects factor. There was a significant effect of problem type on accuracy rates [$F(3, 394) = 26.16, p < .001, \eta^2 = .17$]. Post-hoc comparisons using the Tukey HSD test indicated that the accuracy rate for projection ($M = 63.59, SD = 23.33$) was significantly higher than union ($M = 42.53, SD = 28.23$) ($p < .001$); accuracy of projection ($M = 63.59, SD = 23.33$) was significantly higher than join ($M = 29.08, SD = 21.52$) ($p < .001$); accuracy of projection ($M = 63.59, SD = 23.33$) was significantly higher than difference ($M = 51.88, SD = 41.33$) ($p < .05$); accuracy of union ($M = 42.53, SD = 28.23$) was significantly higher than join ($M = 29.08, SD = 21.52$) ($p < .01$); and accuracy of join ($M = 29.08, SD = 21.52$) was significantly lower than difference ($M = 51.88, SD = 41.33$) ($p < .001$). There was no statistically significant difference between union and difference ($p = .13$).

The one-way ANOVA to compare discrimination rates across different types of problems also revealed a significant effect [$F(3, 394) = 25.47, p < .001, \eta^2 = .16$]. Post-hoc comparisons using the Tukey HSD test showed that the discrimination rate for projection ($M = 62.61, SD = 27.58$) was significantly higher than union ($M = 30.71, SD = 27.01$) ($p < .001$); projection ($M = 62.61, SD = 27.58$) was higher than join ($M = 40.05, SD = 37.10$) ($p < .001$); union ($M = 30.71, SD = 27.01$) was lower than difference ($M = 62.63, SD = 34.70$) ($p < .001$); and join ($M = 40.05, SD = 37.10$) was lower than difference ($M = 62.63, SD = 34.70$) ($p < .001$). There were no statistically significant difference between projection and difference ($p = 1.0$) or union and join ($p = .18$).

Our results suggest that projection is the easiest while join is hardest of all the four problems, given that projection has the highest accuracy and discrimination rates while join has the lowest.

2.4.2. Relationships between familiarity and performance

Table 1 shows the descriptive statistics of the familiarity scores. The results show that the median of the total score ($Mdn = 29$) is lower than the Midrange (= 40). Results also show that the median of Excel Score ($Mdn = 13$) is slightly lower than the median of Rest Score ($Mdn = 15$).

Table 1: Descriptive statistics of experience scores

	Median	Mode	Range	Minimum	Maximum
Total score	29	19	60	10	70
Excel Score	13	6	30	5	35
Rest Score	15	15	30	5	35

In order to test the effects of familiarity (Excel vs. logical) on accuracy and discrimination, we performed correlation analysis between the accuracy and discrimination rates for each problem and the familiarity scores. Table 2 shows the correlations of the accuracy rates with the total familiarity scores (sum of scores for all 10 questions), familiarity in Excel scores and familiarity with logic, math and problem solving (the rest of the questions), respectively. Our results show a positive correlation between the discrimination rates and the familiarity scores (Total, Excel and Rest) for the join problem, suggesting that participants with more familiarity with Excel and other relevant topics (like programming, logic, etc) were better at identifying the incorrect steps. It is interesting to note that none of the other correlations were significant, suggesting that join is the only operation where familiarity correlated with performance.

*Table 2:*Correlations of Familiarity Score (Total Score, Excel Score and Rest Score)and Accuracy and Discrimination Rates

	Total Score				Excel Score				Rest Score			
	Accuracy		Discrimination		Accuracy		Discrimination		Accuracy		Discrimination	
	r_s	N	r_s	N	r_s	N	r_s	N	r_s	N	r_s	N
P	-.04	115	.05	115	-.09	115	-.07	115	-.05	115	.03	115
U	-.10	92	-.10	92	-.02	92	-.11	92	-.17	92	-.10	92
J	-.10	98	.23*	98	-.08	98	.23*	98	-.18	98	-.24*	98
N	.23	93	.03	93	.01	93	.02	93	.03	93	-.11	93

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

2.4.3. Relationships between difficulty and confidence and performance

We also analyzed the self-report scores (difficulty and confidence) of the participants and their relationship to accuracy and discrimination. We posit that if the participants understood the problem and were able to evaluate their performance, then it should be reflected in the correlations between the accuracy rates and self-report. Table 3 shows the results of the correlations between the difficulty/confidence scores with the accuracy rates and discrimination rates respectively. Our results show that for all the problems at least one out of four possible correlations (Accuracy vs. Confidence, Accuracy vs. Difficulty, Discrimination vs. Confidence, and Discrimination vs. Difficulty) is significant. For Projection, all four correlations are significant. On the other hand, for Join three correlations are significant. Union and difference each have one significant correlation. These results suggest that participants have more accurate self-report scores for Projection and Join, which seem to be the simplest and most complex of all the problems. These results indirectly support the hypothesis that some problems are more complex than the others, and participants are able to judge it consciously as determined by their self-report scores.

Table 3: Correlations of Difficulty and Confidence level with Accuracy and Discrimination rates

	Accuracy				Discrimination			
	Difficulty		Confidence		Difficulty		Confidence	
	r_s	N	r_s	N	r_s	N	r_s	N
P	-.25**	115	.40***	115	-.41***	115	.35***	115
U	.18	92	-.06	92	.13	92	.21*	92
J	-.35***	98	.33**	98	-.22*	98	-.08	98
N	.12	93	.36***	93	.08	93	.13	93

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

2.5. Summary

Study 1 results reveal a pattern in problem complexity. As we hypothesized, the results suggest that join is the most complex, and projection is the simplest of all the operations. This result confirms our hypothesis that operations differ in complexity. Looking at the correlations of the familiarity scores and accuracy and discrimination rates, the results are not as expected. We hypothesized that familiarity scores should correlate with the performance in all the relational problems (to different degrees) but our results did not reveal this. Instead, the familiarity was correlated with performance for only one problem: join. This further suggests that relation of problem complexity and familiarity. Given familiarity was only related to performance for join, it shows that familiarity with tabular representations of data, as given by spreadsheets, only helps for more complex relational operations but not for simpler ones. This leads to the question that what happens when these operations are combined into more complex problems? Can familiarity play more prominent in more complex problems?

In study 2, we investigate more complex relational problems that are the combinations of basic level relational operations. Here again we investigate the role of problem complexity and participant's familiarity on their performance, as determined by their accuracy in solving the problems.

3. Study 2

In study 2, we focus on the combinations of elementary operations (projection, union, join and difference) we looked at in study 1. We investigate the role of problem complexity and participant's familiarity with spreadsheets and other relevant topics on their accuracy of solving each type of problem.

3.1. Participants

Participants were recruited from Amazon mTurk for a “Relational Problem Solving” study and were compensated at a flat rate of \$0.50 for participation. No bonus was awarded for performance. A total of 403 participants completed the study. The mean age of the participants was 34.45 ($SD = 12.25$), and 42% of the participants were female. On average, participants took 5 minutes and 28 seconds ($SD = 2$ minutes and 40 seconds) to complete the study.

3.2. Design and Procedure

The experiment design for study 2 was exactly the same as study 1. The only difference was in the type of problems. Participants solved one relational problem, which was based on the

combination of four basic level relational operations. We designed four relational problems: a) projection + selection b) projection + selection + union c) projection + selection + difference d) projection + selection + join. Each problem constituted one condition. Like study 1, the design was a between-subjects experiment where participants were assigned to one of four conditions randomly. After completing the task, participants responded to the same self-report questions as in the study 1 about difficulty and confidence level. At the end, they completed the familiarity questionnaire, which consisted of the same questions as in Study 1.

3.3. Results

3.3.1. Accuracy in the different types of problems

Figure 2 represents the mean accuracy and discrimination rates for each of the four relational problems. As can be seen, problems differ in their mean accuracy rates, suggesting that problems differ in complexity.

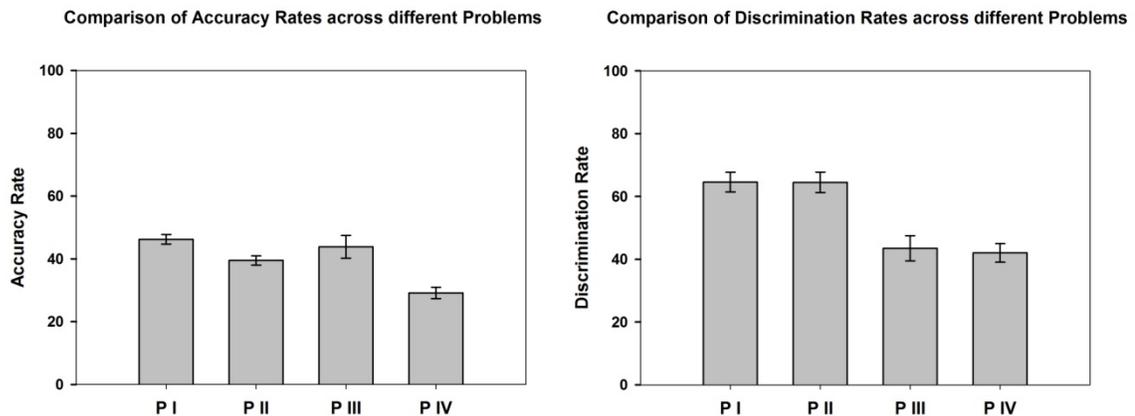


Figure 2: Left panel: Mean accuracy rates for the four different relational problems. Right panel: Mean discrimination rates for the four different relational problems. *Note:* Error bars indicate standard error of the mean.

In order to study the effect of problem type on accuracy and discrimination rates, a one-way ANOVA was performed with problem type (Problem I vs. Problem II vs. Problem III vs. Problem IV) as the between-subjects factor. There was a significant effect of problem type on Accuracy rates [$F(3, 398) = 11.57, p < .001, \eta^2 = .08$]. Post-hoc comparisons using the Tukey HSD test indicated that the Accuracy rate for Problem I ($M = 46.25, SD = 16$) was significantly higher than Problem IV ($M = 29.17, SD = 17.22$) ($p < .001$); accuracy rate for Problem II ($M = 39.49, SD = 15.18$) was significantly higher than Problem IV ($M = 29.17, SD = 17.22$) ($p < .01$), and accuracy rate for Problem III ($M = 43.84, SD = 34.74$) was significantly higher than Problem IV ($M = 29.17, SD = 17.22$) ($p < .001$). There were no statistically significant differences between Problem I and Problem II ($p = .10$), Problem I and Problem III ($p = .86$), and Problem II and Problem III ($p = .50$).

The one-way ANOVA to compare discrimination rates across different types of problems also revealed a significant effect [$F(3, 398) = 14.12, p < .001, \eta^2 = .10$]. Post-hoc comparisons using the Tukey HSD test showed that the discrimination rate for Problem I ($M = 64.56, SD = 33.11$) was significantly higher than Problem III ($M = 43.48, SD = 38.19$) ($p < .001$); discrimination rate of Problem I ($M = 64.56, SD = 33.11$) was significantly higher than Problem IV ($M = 42.03, SD = 28.33$) ($p < .001$); discrimination rate of Problem II ($M = 64.49, SD = 33.48$) was significantly higher than Problem III ($M = 43.48, SD = 38.19$) ($p < .001$), and discrimination rate of Problem II ($M = 64.49, SD = 33.48$) was significantly higher than Problem IV ($M = 42.03, SD = 28.33$) ($p < .001$). There were no statistically significant differences between Problem I and Problem II ($p = 1.0$) and Problem III and Problem IV ($p = .99$).

Our results suggest that Problem I is the easiest while Problem IV is hardest of all the four problems, given Problem I has the highest accuracy and discrimination rates while Problem IV has the lowest.

3.3.2. Relationships between familiarity and performance

Table 4 shows the descriptive statistics of the familiarity scores. Like Study 1, here also we observed that the median of the total score ($Mdn = 29$) was lower than the midrange (= 40). Also, the median of Excel Score ($Mdn = 12$) was slightly lower than the median of Rest Score ($Mdn = 16$).

Table 4: Descriptive statistics of experience scores

	Median	Mode	Range	Minimum	Maximum
Total Score	29	37	58	10	68
Excel Score	12	9	30	5	35
Rest Score	16	17	30	5	35

Table 5 shows the correlations of the accuracy rates with the total familiarity scores, excel familiarity scores and rest of familiarity scores, respectively. Our results show a positive correlation between the accuracy rate and the Excel score for Problem IV, suggesting that participants with more experience with Excel were better at choosing and sequencing the correct steps for join with projection and selection. It is interesting to note that we got a positive correlation between accuracy and familiarity scores for the join problem in Study-1. None of the other correlations were significant, suggesting that Problem IV (join with projection and selection) is the only problem where experience affected the performance.

Table 5: Correlations of familiarity scores (Total Score, Excel Score and Rest Score) with Accuracy Rate and Discrimination Rate

	Total Score				Excel Score				Rest Score			
	Accuracy		Discrimination		Accuracy		Discrimination		Accuracy		Discrimination	
	r _s	N										
P I	.01	111	.005	111	.09	111	-.06	111	.10	111	-.07	111
P II	.05	107	-.07	107	.07	107	-.15	107	.002	107	-.03	107
P III	-.01	92	-.05	92	-.01	92	-.06	92	-.03	92	-.04	92
P IV	.19	92	-.04	92	.25*	92	-.04	92	.06	92	-.03	92

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

3.3.3. Relationships between difficulty and confidence and performance

Similar to study 1, the self-report scores (difficulty and confidence) of the participants were analyzed using correlation analysis of the difficulty and confidence scores with the accuracy and discrimination rates. Following a similar pattern of study 1, our results show that all but one problem has at least one out of four possible correlations significant. The only problem which did not have any of the correlations to be significant was Problem III, which combined difference with projection and selection.

Table 6: Correlations of Difficulty and Confidence level with Accuracy and Discrimination rates

	Accuracy				Discrimination			
	Difficulty		Confidence		Difficulty		Confidence	
	r _s	N						
P I	-.19*	111	.13	111	-.16	111	.26**	111
P II	-.45***	107	.40***	107	.60***	107	.48***	107
P III	.10	92	-.20	92	-.03	92	-.08	92
P IV	.17	92	-.25*	92	.20	92	.26*	92

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

3.4. Summary

Results from Study 2 reveal that problems that are the combinations of different elementary operations also differ in complexity. Here again, we see a similar pattern as study 1: problem which contained join in combination of projection came out to be the most complex, while problem containing projection came out to be the simplest. This was revealed from the accuracy and discrimination rates, which further confirms our hypothesis. The familiarity scores also revealed a similar pattern. The only significant correlation of the familiarity and accuracy was revealed for problem IV, which contained the join operation. The fact that the results from study 2 complement the results from study 1, strengthens our hypothesis, suggesting that relational operations differ in complexity and familiarity only help in solving problems that contain more complex relational operations.

4. Discussion

One of the major motivations of this paper was to study how people do relational reasoning. Previous studies in the field of psychology of human reasoning offer little insight as to how people solve relational problems. In this effort, we presented two studies that looked at human performance in relational problems composed of some of the basic level relational operators. We were interested in knowing the role of problem complexity on performance as determined by accuracy and discrimination rates, and the role of familiarity with spreadsheets and other relevant topics on performance.

Our results show that basic relational operations (projection, union, join, difference) differed in complexity. This was revealed in different accuracy and discrimination rates across problems in both the studies. Our results suggest projection (high accuracy and discrimination rates) was the easiest while join (low accuracy and discrimination rates) was the toughest of all the operations. This not only confirmed our hypothesis that different operations have different complexity; it also showed a pattern of complexity among these operations. We also found that accuracy and discrimination rates followed similar patterns i.e., problems with high accuracy rates tend to have high discrimination rates too. This suggests that subjects who were good at choosing correct options (accuracy) were also good at ignoring incorrect options (discrimination) and vice versa. This result informs us that in an automated assistant, like NEXCEL, more complicated operations like join will need a more intuitive handling. In the currently available spreadsheet applications, there is no direct way to perform a join operation, even though it is a pretty simple concept when described in terms of tables. In our studies, the steps involved to solve a given problem were described in rows and columns terminology, which are more cognitively demanding for an average user. In a deductive spreadsheet, these basic level operations will be defined as the building blocks for more complicated operations. Hence, it will shield the user from the cognitive load of manipulating rows and columns individually.

The fact that different relational operations differ in complexity, as revealed from the performance, implies that different operations utilize different levels of cognitive processing. Simple operations like projection require less cognitive processing to perform, while more complex operations like join require higher cognitive processing. This may include loads on working memory and other cognitive functions. We predict that mental-model analyses (Johnson-Laird, 1991) of relational inference will reveal similar patterns of results as seen in our studies. There are several reasons for this prediction but the most important one is the fact that our experimental results match the implementation complexity of relational operations i.e., operations that are more complex (like join) to implement computationally for a system like NEXCEL are also more difficult to perform manually. Given computational complexity will match the complexity revealed by mental-models approach, we believe that we can make this prediction. We believe that this connection is important in the sense that it will inform the developers of a system like NEXCEL, how to manage complex operations such that they can be intuitive and may reduce the cognitive load of the user. Also, this should motivate future analyses of relational reasoning.

Our analyses of the familiarity score with the accuracy and discrimination rates did not reveal any significant effects for any of the operations but join. Our results suggest that discrimination rates were positively correlated with the familiarity score for join, showing that subjects with more familiarity with the spreadsheet application and other relevant topics

(computer programming, logic, etc) were better at discriminating the incorrect options from the correct options. It is interesting to note that none of the other correlations were significant; suggesting that familiarity (or experience) with spreadsheets did not affect the performance in solving the relational problems. There can be several reasons for this. However, the most relevant one can be the fact that problems other than join were not challenging enough to bring out the difference between subjects with high and low familiarity, and familiarity does not play as significant role in relational reasoning as we hypothesized earlier.

Analyses of self-report measures also revealed interesting results. Our results suggest that self-reports were more accurate for projection and join but not so accurate for union and difference, suggesting that subjects had a better perception of the difficulty of the problem and their confidence in their solution for the easiest (projection) and toughest (join) problems. The correlations of the difficulty and confidence level with the discrimination were similar.

The results from both the studies solidify our claim that different relational operations have different complexity and utilize different cognitive processing. We saw that accuracy and discrimination rates were dependent on problem complexity. We also saw that familiarity with spreadsheets as an instance of tabular data representation did not affect performance in different relational problems. These results give some insight as to how people do relational reasoning. Although we have barely scratched the surface of this field of research, we believe that these findings will motivate future research in relational reasoning.

Some of the questions we did not find answers to include: 1) Do people utilize strategies from everyday reasoning (Oaksford and Chater, 2001) for solving relational problems? 2) If yes, do those strategies relate to the errors they make? 3) What particular types of errors do people make while solving different relational problems? 4) How can we rectify those errors? 5) What can we learn from the patterns of such errors? We believe that these are important questions and any insight they can give us will help us better understand how people do relational reasoning.

In the probabilistic approach to human reasoning, Oaksford and Chater argue that people utilize strategies from everyday reasoning while they solve reasoning tasks in the laboratory (Oaksford and Chater, 2001; Oaksford and Chater, 2007). We wonder if people do the same while they do relational reasoning. We know people use relations in their everyday life. Can we expect them to bring some of those strategies to the laboratory? There is a good reason to believe that people might do that, which might lead to suboptimal solutions or even errors. Future work in this area should account for such issues.

The goal of this paper was to study how humans solve relational problems. Given relational problems have only been studied from the perspective of computer science; here we have presented it from a psychological perspective. We hope that this effort will lead to further research in HCI and other disciplines that relate to cognitive aspects of relational problem solving. Further research is required to make insightful connection between relational calculus and human reasoning.

References

- Cervesato, I. (2007). NEXCEL, a deductive spreadsheet. *The Knowledge Engineering Review*, 22(03), 221-236.
- Cervesato, I. (2013). *The deductive spreadsheet* (pp. 183-252). Springer Berlin Heidelberg.
- Chan, D. and Chua, F. (1994) Suppression of valid inferences: syntactic views, mental models, and relative salience. *Cognition* 53, 217–238
- Evans, J.St-B.T. *et al.* (1993) *Human Reasoning*, Erlbaum
- George, C. (1997) Reasoning from uncertain premises. *Think. Reason.* 3, 161–190
- George, C. (1999) Evaluation of the plausibility of a conclusion from several arguments with uncertain premises. *Think. Reason.* 5, 245–281
- Johnson-Laird, P. N., & Byrne, R. M. (1991). *Deduction*. Lawrence Erlbaum Associates, Inc.
- Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological review*, 99(3), 418.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.
- Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press.
- Liu, I. M. (1996). A Probabilistic Interpretation of If-Then. *The Quarterly Journal of Experimental Psychology: Section A*, 49(3), 828-844.
- Manktelow, K. (1999) *Reasoning and Thinking*, Psychology Press
- Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(4), 883.
- Oaksford, M. and Chater, N., eds (1998) *Rational Models of Cognition*, Oxford University Press
- Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in cognitive sciences*, 5(8), 349-357.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
- Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32(01), 69-84.
- Rips, L.J. (1994) *The Psychology of Proof*, MIT Press

Stein, E. (1996) *Without Good Reason*, Oxford University Press

Schroyens, W., Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2000). Conditional reasoning with negations: Implicit and explicit affirmation or denial and the role of contrast classes. *Thinking & reasoning*, 6(3), 221-251.

Schroyens, W., Schaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in propositional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6), 1713.

Stevenson, R.J. and Over, D.E. (1995) Deduction from uncertain premises. *Q. J. Exp. Psychol. A* 48, 613–643

Stich, S. (1985) Could man be an irrational animal? *Synthese* 64, 115–135