Limited Domain Structure for Conjunction Errors

Ethan Ludwin-Peery[1] (elp327@nyu.edu)
[1]Department of Psychology, NYU, New York

Abstract

People make conjunction errors, rating a conjunction as more likely than one of its constituents, across many different types of problems. They commit the conjunction fallacy in problems of social judgment, in physical reasoning tasks, and in gambles of pure chance. Doctors commit the fallacy when making judgments about hypothetical patients. Do all these errors share an underlying cause? Or does the fallacy arise independently in different types of reasoning? In a series of studies, we look for structure in conjunction errors across various types of problems. We find that error magnitudes are related for some clusters of items, but there does not appear to be a universal relationship between all cases of this fallacy.

Keywords: fallacies; heuristics; rationality; conjunction fallacy

Introduction

Linda — "deeply concerned with issues of discrimination and social justice" — sounds very much like one's stereotype of a feminist, but not so much like one's stereotype of a bank teller. As a result, she seems more likely to be both a feminist and a bank teller than a bank teller in general, even though this is logically impossible. In their original work on the conjunction fallacy, Tversky and Kahneman (1980, 1982) argue that conjunction errors arise from this kind of stereotype-based reasoning, which they identify as the representativeness heuristic.

Recent work has attempted to specify computational theories that can more precisely explain how such conjunction errors come about, and a diverse set of perspectives have been advanced. Some suggest that conjunction errors arise from the introduction of noise to the measurement of the constituent probabilities (Costello, 2009); Bayesian accounts indicate that sampling from vast data spaces naturally gives rise to classic reasoning errors, including the conjunction fallacy (Sanborn & Chater, 2016); there are even theories which propose that conjunction errors necessarily follow when quantum probability is used rather than classical probability (Busemeyer, Pothos, Franco, & Trueblood, 2011; Pothos & Busemeyer, 2013).

All of these accounts, however, share the implicit assumption that conjunction errors in different tasks arise from the same cognitive mechanism, but this doesn't necessarily have to be the case. The term "conjunction error" refers to an observed behavior, any case where a conjunction ($A \wedge B$) is rated as more likely than one of its constituents (either A or B, or very rarely both). But in the same way that a traffic accident might be caused by fatigue, distraction, or the driver having a stroke, this particular error could have multiple potential causes.

Conjunction errors have been observed across a surprising variety of problems; not only judgments of stereotypes, as in the famous Linda problem, but also in judgments of medical conditions by physicians, estimates of Wimbledon victory in the year 1981, and evaluations of various possible outcomes when rolling colored dice (Tversky & Kahneman, 1983). Different mechanisms might be recruited in response to these very different problems, and if so, the errors may not share an underlying psychological cause. Even if similar processes are used to estimate probability in all cases, there might be multiple points of failure that could each independently cause conjunction errors.

It's rare, but sometimes we discover that apparently singular phenomena are not so closely related after all. Theory of mind, for example, has traditionally been considered to be a single construct, but recent work has found that different tasks intended to measure this construct show minimal correlations with one another (Warnell & Redcay, 2019). Something similar could be the case for conjunction errors. There may not be a conjunction fallacy *per se*; it might instead be the case that conjunction errors occur as the result of different cognitive mishaps in different situations.

Study 1A

Previous work has used patterns of correlations to argue that performance on different tasks or measures of ability are or are not cognitively related. For example, Dillon, Huang, and Spelke (2013) showed that some forms of geometric reasoning were related in children, while others were not, making a convincing case that children are making use of at least two different types of geometric representations. Warnell and Redcay (2019) found evidence that some theory of mind tasks—such as various false belief tasks—were related, while many other measures were not related, suggesting that theory of mind is a useful concept, but may be more multidimensional than previously understood. A similar design is appropriate for questions involving conjunction errors.

To behaviorally measure whether varied conjunction fallacy questions share an underlying structure, we selected a range of questions from previous research. As the first goal

was to determine whether there is structure in this domain at all, we started with a diverse set of tasks chosen to cover a variety of domains and modalities. While it is not possible to know how to select maximally dissimilar materials, as we don't know what dimensions are relevant to the structure (if any) of the fallacy, the materials were selected with maximum diversity in mind.

While the questions themselves were chosen for diversity, in every case, participants were asked to estimate likelihoods in percent chance. Holding the method of judgment constant across all questions is important because variations in the form of the question — for example, asking for frequency rather than probability judgments — has been found to influence the rate of conjunction errors (Fiedler, 1988).

In study 1A, we compared six problems that have been reported to reliably lead to conjunction errors.

Participants

We collected 209 participants from Amazon Mechanical Turk (78 women; M_{age} = 33.47; SD_{age} = 9.62). Participants were excluded if they did not not complete the study or if they failed certain comprehension questions. All exclusion criteria were preregistered.[1] In total, 105 passed all exclusion criteria (44 women; M_{age} = 34.85; SD_{age} = 10.40).

We analyzed only the first 100 participants (41 women; M_{age} = 35.00; SD_{age} = 10.54), as preregistered.

Methods

The first item was a physical conjunction fallacy task first reported by Ludwin-Peery, Bramley, Davis, and Gureckis (2019). In this task, participants view the first few moments of several scenes and rate the probability of a future event occurring (e.g., what is the probability that the ball will fall into the hole?). Each of eight critical scenes appears twice, once with a conjunction question and once with a constituent, and the key dependent variable is the difference of the two estimated probabilities. Because the task involves a cannonball and a sphere, to distinguish it from other physical reasoning tasks that might elicit conjunction errors, we call it the **Cannonball & Sphere item**, or **C&S** for short.

The second item was adapted from Sides, Osherson, Bonini, and Viale (2002). We called this problem **Taxes**. The remaining four problems were taken from Tversky and Kahneman (1983). These questions historically have elicited the conjunction fallacy in different domains. We called them **Bill**, **Peter**, **Health**, and **Dice** for short. The full text and materials for all problems is available on the OSF.

Results

For the C&S item we averaged the rating difference scores (% sole probability - % conjunction probability) of each participant for each of the eight C&S scenes. We calculated the magnitude of the conjunction errors for the rest of the items by calculating a difference score (% sole probability - % conjunction probability) for the two critical judgments.

[1] Preregistration form here.

Table 1: Pearson Correlations Among Conjunction Errors in Study 1A

	C&S	Taxes	Bill	Peter	Health	Dice
C&S	–					
Taxes	-0.099	–				
Bill	0.006	-0.153	–			
Peter	-0.057	0.194	-0.023	–		
Health	0.022	0.120	-0.032	0.193	–	
Dice	0.201*	-0.310**	0.202*	-0.064	0.175	–

*, unadjusted $p < .05$
**, unadjusted $p < .01$

Conjunction Fallacy The first question for all items was whether they had actually elicited conjunction errors.

Two-tailed one-sample *t*-tests found that participants consistently rated the conjunction outcomes as more likely than their constituents for the **Cannonball & Sphere** question, $t(99) = 5.63$, $p < .001$, 95% confidence interval of the difference [4.75, 9.93], the **Taxes** question, $t(99) = 2.17$, $p = .032$, 95% confidence interval of the difference [0.61, 13.43], the **Bill** question, $t(99) = 4.49$, $p < .001$, 95% confidence interval of the difference [5.83, 15.09], the **Health** question, $t(99) = 2.64$, $p = .009$, 95% confidence interval of the difference [1.31, 9.21], and the **Dice** question, $t(99) = 4.08$, $p < .001$, 95% confidence interval of the difference [4.75, 13.75].

For the **Peter** question, participants actually rated the constituent outcome as *more* likely than the conjunction, $t(99) = -3.43$, $p < .001$, 95% confidence interval of the difference [-11.58, -3.10]. This is logically sound and therefore not an example of the conjunction fallacy.

Correlations In order to account for multiple comparisons, we used a Bonferroni-corrected alpha, as preregistered, .05 / 15 = .00333 for our new alpha.

As seen in Table 1, the magnitude of correlation across the various conjunction fallacy problems was quite small.

The correlation between the error magnitudes for **Dice** and **Taxes** was significant even with our corrected alpha, $r(98) = -0.31$, $p = .002$. But surprisingly, this correlation was negative, suggesting that people who make more extreme conjunction errors on the **Dice** problem actually make *less* extreme conjunction errors on the **Taxes** problem. This is not what we would expect if these errors had a common cause.

It appears that the magnitude of conjunction errors across different tasks is not reliably related. Individuals who make large errors on one question do not seem to be more likely to make similarly extreme errors on another question.

Exploratory Factor Analysis In an exploratory factor analysis, all eigenvalues were less than 1, strongly suggesting no factor structure.

Discrete Conjunction Relationships Conjunction fallacy errors can be measured in magnitude, but we can also simply measure the presence or absence of the fallacy. If someone rates a conjunction as more likely than its constituent, then they have committed the conjunction fallacy regardless of how large the difference is.

We converted the conjunction errors from all our items to Boolean variables, where any value greater than zero (indicating that they rated the conjunction as more likely than its

Table 2: Chi-Square Tests of Relation Among Conjunction Errors in Study 1A

	C&S	Taxes	Bill	Peter	Health	Dice
C&S	–					
Taxes	0.007	–				
Bill	0.011	0.988	–			
Peter	0.000	2.459	0.044	–		
Health	1.132	12.971 ***	1.408	1.083	–	
Dice	0.347	0.178	6.253 *	0.096	0.399	–

*, unadjusted $p < .05$
***, unadjusted $p < .001$

constituent) was treated as a True. We then conducted Chi-squared tests of association for each of the pairs of items. The result of these tests are reported in Table 2.

Somewhat surprisingly, these analyses paint a very different picture of the relationships between the items. Here, taxes and health are most closely related items, and significant with correction for multiple comparisons (even if we use a Bonferroni correction for 30 tests instead of 15). The second most closely related pair is Bill and Dice, though this is not significant with correction.

This analysis is particularly interesting in how closely it matches the pattern observed in the factor analysis we perform in Study 2 (see below).

Discussion

This result, finding almost no evidence of a category structure, was unusual and unexpected. To confirm this finding, we decided to do a direct replication on a different population.

Study 1B

In Study 1B, we ran a direct replication of study 1A on a population of undergraduate students.

Participants

We collected 166 participants from New York University's student subject pool (105 women; M_{age} = 19.44; SD_{age} = 1.49).

Of those, exactly 100 passed all exclusion criteria (67 women; M_{age} = 19.58; SD_{age} = 1.49), and we analyzed only the first 100, as preregistered.[2]

Methods

All methods were identical to the methods used in Study 1A.

Results

We calculated conjunction errors in the same manner as in Study 1A.

Conjunction Fallacy Again, the first question to ask was whether these problems actually elicited conjunction errors.

Two-tailed one-sample *t*-tests found that participants consistently rated the conjunction outcomes as more likely than their constituents for the **C&S** task, $t(99) = 4.72$, $p < .001$, 95% c, the **Bill** question, $t(99) = 4.04$, $p < .001$, 95% confidence interval of the difference [4.53, 13.28], the **Health** question, $t(99) = 4.32$, $p < .001$, 95% confidence interval of the difference [4.73, 12.75], and the **Dice** question, $t(99) = 3.64$, $p < .001$, 95% confidence interval of the difference [2.12, 7.21].

[2]Preregistration form here.

Table 3: Pearson Correlations Among Conjunction Errors in Study 1B

	C&S	Taxes	Bill	Peter	Health	Dice
C&S	–					
Taxes	-0.031	–				
Bill	0.106	0.055	–			
Peter	-0.184	0.202*	-0.122	–		
Health	0.138	-0.096	0.086	-0.051	–	
Dice	-0.014	-0.031	0.154	0.031	0.031	–

*, unadjusted $p < .05$

Table 4: Chi-Square Tests of Relation Among Conjunction Errors in Study 1B

	C&S	Taxes	Bill	Peter	Health	Dice
C&S	–					
Taxes	0.017	–				
Bill	0.079	1.217	–			
Peter	1.145	4.043 *	0.510	–		
Health	0.98	1.411	6.648 *	0.805	–	
Dice	3.428	3.143	1.612	0.409	0.000	–

*, unadjusted $p < .05$

As in Study 1A, a two-tailed one-sample *t*-test found that for the **Peter** question, participants actually rated the constituent outcome as *more* likely than the conjunction. Unlike in Study 1A, in this sample, a two-tailed one-sample *t*-test found no difference for the **Taxes** question, $p = .722$.

Correlations As before, in order to account for multiple comparisons, we used a Bonferroni-corrected alpha, as preregistered.

All correlations from Study 1B are presented in Table 3. Only one of these correlations was of notable magnitude, the correlation between the **Peter** and the **Taxes** problems, but it was not significant with our corrected alpha.

Combination with Study 1A When these data are pooled with the data from study 1A (total $n = 200$), three correlations are significant at $p < .01$, but none have *p*-values less than the Bonferroni-corrected alpha of .00333.

Discrete Conjunction Relationships As before, we also looked at the relationship between error *commission* among items, as shown in Table 4. Unlike in Study 1A, in this case, little about the analysis changes. None of the items show relationships that are significant after correction.

Study 2

So far we have not found much evidence for a relationship between the magnitude of the errors from different questions that elicit conjunction errors. At this point, we want to know if this is evidence that there is no relationship to be found, or if it means that we are simply not searching in the right way.

Noise Perhaps it is simply not possible to find these relationships, even if all conjunction errors come from the same cognitive mechanism. If the particular magnitude (rather than the direction) of the error were simply random noise, then the conjunction errors would never be correlated.

If this were the case, then the magnitude of conjunction errors would not be correlated even for conjunction fallacy

problems that very closely resemble one another.

Power It's also possible that the first studies were underpowered. The true correlations might be real, but quite small; perhaps the sample size was too small to detect them.

A sample size of 100 has about 80% power to find a correlation of 0.27 and about 90% power to find a correlation of 0.32. Our total sample size of 200 has about 80% power to find a correlation of 0.19 and about 90% power to find a correlation of 0.23. The problem is that we simply don't know what magnitude of correlation to expect.

One way to deal with both these issues is to test intentionally similar items. By comparing very similar problems, we can estimate a baseline of how correlated these errors can be.

Participants

We collected 332 participants from New York University's student subject pool (198 women; M_{age} = 19.39; SD_{age} = 1.31). Of those, exactly 200 passed all exclusion criteria (141 women; M_{age} = 19.40; SD_{age} = 1.25). Exclusion criteria were the same as in previous studies. We analyzed only the first 200, as preregistered.[3]

Methods

The C&S task is already the result of the aggregation of multiple pairs of questions, and was left unchanged. We dropped the Peter problem because in Studies 1A and 1B, participants didn't commit the conjunction fallacy on this item.

New Questions For each of the remaining questions, we found two new questions that were intended to closely match the original both in content and in structure. In some cases we drew the new questions from the literature. For example, one of the new questions to match Bill is the infamous Linda problem (Tversky & Kahneman, 1982). In other cases we modified the original questions or developed new questions from scratch. To conserve space, the full text and materials for the new problems is available on the OSF.

Results

Conjunction Fallacy As before, the first question to ask is whether these problems actually elicited conjunction errors. Errors were calculated in the same way as in previous studies.

A a two-tailed one-sample t-test found that for the **C&S** task, participants consistently rated the conjunction outcomes as more likely than their constituents, $t(199) = 6.40$, $p < .001$, 95% confidence interval of the difference [4.33, 8.19].

Two-tailed one-sample t-tests found that **Taxes 1** did not cause reliable conjunction errors, $p = .399$. However, the other two **Taxes** questions produced reliable errors, all t's > 4.0. Two-tailed one-sample t-tests found that all three **Bill** questions caused reliable conjunction errors, all t's > 4.5. Two-tailed one-sample t-tests found that all three **Health** questions caused reliable conjunction errors, all t's > 3.0. Two-tailed one-sample t-tests found that all three **Dice** questions caused reliable conjunction errors, all t's > 3.0.

[3] Preregistration form here.

Correlations In order to account for multiple comparisons, we again used a Bonferroni-corrected alpha, as preregistered. With 13 items, there are a total of 78 possible 2-pair combinations, .05/78 = .00064 for our new alpha.

This new alpha is quite small, but a sample size of 193 has 80% power to detect a correlation of moderate size (r = 0.3), even with this adjusted alpha. As our total sample size is 200 participants, this study is reasonably powered to detect moderate correlations between these errors, should they exist.

The full set of correlations appears in Table 5, and scatterplots of all comparisons are available on the OSF. The strongest relationship observed, between Health 1 and Health 2, had a correlation coefficient of .71. In total, 15 of the 78 possible correlations were significant at the Bonferroni-corrected alpha of 0.00064. A reviewer noted that Bonferroni can be overly conservative, and so we also corrected these correlation tests with False Discovery Rate. With this correction, 30 of the 78 possible correlations were significant. Both criteria are indicated on Table 5.

Many of the observed correlations were between items in the same "family" (for example, Taxes 1 and Taxes 2 were correlated, $r(198) = 0.37$), but some correlations were between items from different families. Health 1 and Taxes 2, for example, were also moderately correlated, $r(198) = 0.43$.

Discrete Conjunction Relationships As before, we also looked at the relationship between error *commission* among items. The results (unfortunately omitted for space concerns) differ in some ways but overall show a similar pattern of relationships.

Exploratory Factor Analysis Initial eigenvalues showed strong support for at least one factor and possible support for a second factor. The first factor explained 19% of the variance and the second 9% of the variance. As a result, we decided to explore both possibilities. One- and two-factor solutions are presented in Table 6.

All thirteen questions were factor analyzed using Promax rotation. We also tested and examined several other rotations (including Varimax and Oblimin), and found almost no difference between solutions using different rotations.

The one-factor solution explains a total of 19% of the variance. Eight items load on this factor with loadings of absolute value .30 or greater, most of them coming from the Taxes and the Heath questions. While the one-factor solution explains a large amount of the variance, it does not seem to explain the overall pattern of conjunction errors. Many of the items do not load strongly onto this factor, including the items that produce the strongest conjunction errors. For example, Bill 2, which is the classic "Linda Problem", has a Cohen's d of .89 but a factor loading of only .21. The C&S item is quite reliable but has a slightly negative loading of -0.08.

The two-factor solution explains a total of 29% of the variance, with the first factor explaining 17% and the second explaining 9%. The two factors are correlated at r = .55, suggesting a moderate relationship between them. Five items load on each factor with loadings of absolute value .30 or

Table 5: Pearson Correlations Among Conjunction Errors in Study 2

	C&S	Taxes 1	Taxes 2	Taxes 3	Dice 1	Dice 2	Dice 3	Health 1	Health 2	Health 3	Bill 1	Bill 2	Bill 3
C&S	-												
Taxes 1	-0.049	-											
Taxes 2	-0.116	0.370 † ‡	-										
Taxes 3	0.008	0.029	0.169 ‡	-									
Dice 1	-0.011	0.137	0.098	0.155	-								
Dice 2	-0.019	0.118	0.145	0.057	0.207 ‡	-							
Dice 3	-0.056	-0.034	0.092	0.058	0.061	0.242 † ‡	-						
Health 1	-0.017	0.244 † ‡	0.431 † ‡	0.335 † ‡	0.251 † ‡	0.150	0.113	-					
Health 2	-0.014	0.188 ‡	0.402 † ‡	0.328 † ‡	0.244 † ‡	0.078	0.177 ‡	0.706 † ‡	-				
Health 3	-0.182 ‡	0.071	0.224 ‡	0.365 † ‡	0.120	0.188 ‡	0.186 ‡	0.445 † ‡	0.475 † ‡	-			
Bill 1	0.035	0.125	0.167 ‡	0.228 ‡	0.111	0.328 † ‡	0.198 ‡	0.237 ‡	0.198 ‡	0.222 ‡	-		
Bill 2	0.089	0.069	0.080	0.111	-0.015	0.134	0.148	0.109	0.077	0.131	0.260 † ‡	-	
Bill 3	-0.062	0.109	0.050	0.096	0.132	0.023	0.089	0.035	0.024	0.056	0.226 ‡	0.141	-

†, unadjusted $p < 0.00064$; ‡, significant according to False Discovery Rate threshold

Table 6: Factor Analysis of Conjunction Errors in Study 2

One-Factor Solution		**Two-Factor Solution**		
	Factor 1		Factor 1	Factor 2
C&S	-	**C&S**	-	-
Taxes 1	0.30	**Taxes 1**	-	-
Taxes 2	0.51	**Taxes 2**	0.48	-
Taxes 3	0.44	**Taxes 3**	0.34	-
Dice 1	0.31	**Dice 1**	-	-
Dice 2	-	**Dice 2**	-	0.50
Dice 3	-	**Dice 3**	-	0.36
Health 1	0.78	**Health 1**	0.87	-
Health 2	0.75	**Health 2**	0.89	-
Health 3	0.58	**Health 3**	0.47	-
Bill 1	0.40	**Bill 1**	-	0.73
Bill 2	-	**Bill 2**	-	0.42
Bill 3	-	**Bill 3**	-	0.32

Rotation Method: Promax
Factor loadings of absolute value less than 0.30 not shown.

greater, and both have items that load strongly. Despite the correlation, there are no cross-loadings. Again, they seem to separate out by question type; most of the Taxes and the Health questions load onto factor 1, as before, and most of the Bill and the Dice questions load onto factor 2.

Some items are still not captured by this solution. If a three- or four-factor solution is fit to the data, the third factor pulls out Taxes 1 without seriously affecting the structure of the first two factors, and the fourth factor isolates the Cannonball & Sphere item into a factor entirely its own. The C&S item's loading on this fourth factor is greater than one, and all other loadings on this factor have absolute values of .15 or less.

Notably, the Taxes and the Dice questions do not separate out onto their *own* factors as more factors are allowed, suggesting that they really do have some commonality with the Health and Bill questions (respectively).

Discussion

Unlike in Studies 1A and 1B, here we found strong and highly significant correlations between the conjunction errors elicited by several items. This clearly establishes that the correlation between such errors can be as strong as $r = .71$, at least when the questions are somewhat similar. Even for relatively dissimilar items, we still saw correlations as strong as $r = .43$. Given the previous results, this was quite surprising.

Ratings of all conjunction fallacy problems were in percentage chance, so the factor structure cannot be the result of the use of different measures between different questions. Previous work has used proportion estimation, rank ordering, and other measures, but percentage chance judgment was held constant here as the method of evaluation.

Rather than simply being correlated with their immediate fellows, items exhibited reliable correlations with items from other groups. This suggests that there is some structure to the commission of these errors, but that the errors are not uniformly committed across all problems that elicit such errors.

Standout Items There appears to be strong support for a 2-factor solution, but there are some refractory items. Why do Taxes 1 and Dice 1 not load cleanly onto their group factors?

Taxes 1 is a question about a tax cut passing Congress, possibly supported by Republicans. It's notable that this study (and Study 1B) used an NYU undergraduate population, which is highly international. International students may not have many stereotypes about the tax policy of the Republican party. If participants don't know that this is a behavior stereotypical of Republicans, they're very unlikely to commit the conjunction fallacy on this problem.

Dice 1 seems to fit more clearly with the Taxes/Health cluster, as evidenced both by the 1-factor solution and by the correlations. The only clear difference is that Dice 1 involves a physical sampling process with replacement (rolling a die), while Dice 2 and 3 involve a psephic sampling process without replacement (drawing jellybeans from a bag or cards from a deck). Could this explain the difference? It's hard to say. Surprisingly, Dice 1 does not seem to be at all correlated with Cannonball & Sphere, which is a strike against the idea that Dice 1 might separate out because it involves some sort of simulation of the roll of the die.

General Discussion

We appear not to have found evidence of any structure in Studies 1A and 1B because, for various reasons, Taxes 1 and Dice 1 do not seem to be especially good examples "of their class". When more items were tested, however, we found evidence of some structure in the errors.

This structure does not appear to be simple. Initially we

expected that if there were clusters, they would likely be thematic in some way. The clusters might for example be related to the content of the questions, or to their structure. As far as we can tell, however, this is not the case.

Representative Conjunctions Tversky and Kahneman (1983) preferred an explanation of the conjunction fallacy in terms of representativeness. Linda, for example, sounds very much like the stereotype of a feminist, and so "Linda is a bank teller and is active in the feminist movement." seems like a decent portrait of her, even though it cannot be a more likely option than "Linda is a bank teller." This explanation fits surprisingly well with the items in the first factor. Note that BGGBGGG seems like a representative sequence of jellybean draws, even though it cannot be more likely than the less obviously representative GGBGGG.

Though not particularly connected in structure or in content, these two groups of items might be connected by a shared aspect of representativeness in their design. In both types of question, the conjunction option has a sort of "at first blush" appeal, where it immediately seems representative of the story described in the question.

This account is somewhat reminiscent of the difficulties presented by the cognitive reflection test (Frederick, 2005), where there is an attractive "intuitive" answer that happens to be entirely incorrect. We might expect CRT scores to correlate with conjunction errors from this group of questions, but maybe not with conjunction errors from other questions.

Narratively Bound Conjunctions The Taxes and Health questions, however, do not have this sense of representativeness. When thinking of events that might happen in the next year, "A renewable energy bill will be passed by Congress between September 1st and December 1st, 2021, and it will be supported by Democrats." does not sound particularly likely. Similarly, "has had at least one stroke and is over 60 years old" does not sound likely to be the description of a randomly selected participant in a survey of general health conducted in California.

The items in the second factor do seem to have something in common, however. What makes these conjunctions enticing is instead their rich logical, possibly causal, structure. A renewable energy bill is likely to be supported by Democrats, and not particularly likely to exist without their support. A randomly selected Californian isn't likely to have had at least one stroke, but it's much more likely if he's over 60. This type of narratively-convincing conjunction may be the feature that ties the Taxes and the Health items together.

Some people may be more likely to see a highly representative case and think of it as very likely, while others are less attracted by the fact that it happens to fit a particular stereotype. And some people may be more likely to judge a story to be highly likely if all the pieces form a coherent, logically supportive narrative. Assuming that there is individual variation and that these two tendencies are not closely associated, this would explain the majority of our results. If this is the case, further research might be able to investigate these two tendencies as possible discrete stages, functions, or strategies used in commonsense reasoning.

Under such an account, novel problems that focus on the similarity of a certain case to a stereotype should fall together with the Dice and Bill items, while novel problems that include causally supportive conjunctions should fall together with the Taxes and Health items. Future work can generate several such items and test this hypothesis.

Cannonball & Sphere Task There is one major problem with this hypothesis, however. The C&S task uses the conjunction, "What is the probability that the cannonball will hit the pink sphere, and then the pink sphere will end up on the grass?" This seems like it is an example of the supportive narrative-style conjunction described above, because the pink sphere seems particularly likely to end up on the grass if the cannonball hits it.

Because it shares this trait, one might expect that this item would fall together with the Taxes and Health items, but it does not. Its largest correlation by magnitude is with Health 3, but the correlation is negative, r = -.18. Its largest positive correlation is only r = .09, with the item Bill 2. And of course, neither of these relationships is significant.

This apparent null correlation is a partial strike against this interpretation. But the C&S task is very unlike the other questions in this survey, unlike them in a number of ways.

First, it involves questions about physical events, and it's possible that reasoning about physical events is qualitatively different from reasoning about more abstract scenarios like the ones described in the other problems. While the Dice problems are in a sense about physical events, participants might equally model them abstractly as problems about sampling with or without replacement, without any consideration towards their physical instantiation. They might equally have done neither. You don't need to run a physical simulation to tell me that it's unlikely to roll a 6 one hundred times in a row, nor do you need to think about the sampling distribution.

Second, it uses video materials, rather than describing a scenario in text. It's possible that richer materials are evaluated differently, leading to similar errors as made when using text, but for different reasons. If we included a version of the Linda problem where, instead of a written description of an outspoken philosophy major, participants were shown a photo of a college-educated 31-year-old from Western Massachusetts, would conjunction errors from that question correlate with the original Linda problem, or with the C&S task?

The evidence so far does seem to somewhat support the idea that the reasoning process used to answer this problem is different from the process used to answer the other problems, despite also leading to conjunction errors. If novel items based on representativeness and causal support fall out as described above, that further supports the idea that something distinct is occurring here, and further work can attempt to disentangle just what it is that makes this task different. In this case, we would expect such investigations to have serious implications for theories of visual and/or physical reasoning.

Acknowledgments The author thanks Maya Goldberg for help with this research, and Adam Mastroianni, Harrison Ritz, and Todd Gureckis for helpful comments and discussion.

References

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193.

Costello, F. J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, *22*(3), 213–234.

Dillon, M. R., Huang, Y., & Spelke, E. S. (2013). Core foundations of abstract geometry. *Proceedings of the National Academy of Sciences*, *110*(35), 14191–14195.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*(2), 123–129.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Ludwin-Peery, E., Bramley, N., Davis, E., & Gureckis, T. (2019). Limits on the use of simulation in physical reasoning. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*(3), 255–274.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*(2), 191–198.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in Social Psychology*, *1*, 49–72.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*, 103997.