Variations in Probability Judgment

Brianna Smith

Department of Psychology

Tufts University

Acknowledgements: I would like to thank Shelby Bean for her work as a research assistant, and especially her help with data collection. I would also like to thank Dan Barch for his assistance with experiment design and advice on collating results.

I. Abstract

This research is a study of accuracy in probability judgment. Previous research in probability and decision making has largely focused on complex social problems with many extraneous factors. In this study, participants were instead given a simple 'toy' problem that reflected pure probabilities. In two experiments, participants were briefly shown a field of marbles and then asked to make judgments about the probability of selecting a certain type of marble from that field. Participants made far more accurate judgments in this task than has been typically found in the existing social-based probability literature. Participants committed very few decision-making fallacies, and their accuracy was neither significantly affected by more difficult presentation modes nor strongly impacted by delays between presentation and test. This finding shows support for the representativeness theory of the conjunction fallacy in decision making. Meanwhile, this result undermines the arguments of other theorists who have assumed that fallacies are due to an essential human difficulty with probability or probability integration.

II. Literature Review

Research in probability judgment has been largely unanimous in finding that participants have difficulty making accurate judgments about probability. Participants frequently commit a range of errors that appear to be logically inconsistent. In the following literature review, I analyze the existing research on some of those errors, focusing on the conjunction and disjunction fallacies.

Probability judgment

Probability judgments are made by everyone, every day. Given the uncertainty of life, we are often called upon to make decisions where we don't know certain factors, frequency rates, or outcomes. We are required to make probability judgments about all of these things in order to function at a reasonable rate in a complicated world. For instance, a probability judgment is required before deciding whether to take an umbrella outside on a cloudy day – will it rain or will it stay dry? Another type of probability judgment is required in almost every social task, as it is impossible to know everything about another person – will someone support you at a project meeting, or will they oppose you? How likely is it? We all must rely on probability judgments in order to continue with our lives. Unfortunately, research has shown that we are also relatively "bad" at probability judgments, in that we tend to deviate from normative models of probability in logically inconsistent ways.

In a series of papers beginning in the 1970s, Daniel Kahneman and Amos Tversky found that participants in their studies made a number of systematically inaccurate probability judgments. The two researchers concluded that the problem was that participants were using heuristics to make judgments, rather than mathematics or logic.

That is, people were using their learned or innate biases about situations in order to guide their probability judgments. This is most clearly seen in the influence of availability on probability judgments. Participants tended to overestimate both the frequency and the probability of an item occurring when those items are easy to retrieve from memory, while underestimating the frequency and probability for items that are difficult to retrieve from memory (Tversky and Kahneman, 1973). For instance, when presented with a list of famous and non-famous names, participants later judged that the frequency of famous and therefore memorable names on the list was higher than the frequency of non-famous names. In fact, there were actually fewer famous names on the list than there were non-famous names. This shows a connection between memory and the ease of retrieval, and heuristics and biases. Tversky and Kahneman argued that participants were using the availability heuristic: participants believed that if they could recall a group of names more easily, then there must have been more of that type of name on the list.

Kahneman and Tversky (1973) showed that participants also tend to ignore base rates or denominators – how often an event occurs within the population as a whole. Instead, participants rely almost entirely on individuating information and predictors given for the case they are supposed to make a probability judgment about. For instance, when given a biographical sketch of a person and asked to make a probability judgment about that person's area of study, participants assigned very high probabilities to what the person 'seemed like,' while ignoring that certain areas of study are far more popular on average than others. This has been termed base-rate neglect, as well as part of the representativeness heuristic, where participants judge something as more likely because it seems to 'fit' the person or situation better. Participants have also demonstrated base-rate neglect in operant paradigms, where they tend to rely more on predictors than the actual base likelihood of reinforcement (Kutzner, Freytag, Vogel, and Fiedler, 2008). Participants also tend to weight predictive information as more important than base rates both when information about base rates is presented numerically and when base rates are presented naturalistically through repeated trials (Stolarz-Fantino, Fantino and Van Borst, 2006; Fantino and Stolarz-Fantino, 2007).

Following the main focus of Kahneman and Tversky's research on the effect of socially-related heuristics, a large segment of the succeeding research in probability judgments since then has largely focused on how and why people perform poorly in socially-relevant probability judgments. Further research has found that participants often discriminate poorly between information that can influence probability judgment, allowing irrelevant information to raise or lower the probability they assign to an event (Dougherty and Sprenger, 2006). This means that the context in which a probability judgment is demanded can matter a great deal, even if the context is irrelevant to the decision. Hanita, Gavanski and Fazio (1997) found that the order in which sample spaces and examples are given can strongly influence probability judgments for conjunctions (the probability of both of two events occurring), even though the base rates for the conjunction's two component events are not affected. This suggests that context and order effects are more important for some kinds of probability judgments than for others.

The general trend of poor ability to make accurate probability judgments can also be seen in studies about cumulative probability. Doyle (1997) found that participants could not accurately judge the probability of an event occurring over a period of several years. Given the likelihood of a contraceptive failing at any given use, participants were asked to estimate the probability of a failed contraceptive over a series of sexual encounters. Most participants used incorrect strategies to arrive at their judgments. One strategy was to keep the failure rate at a constant percentage, rather than accounting for the increasing hazard function of the contraceptive failing given that it had not yet failed. Another incorrect strategy was to multiply the chance of failure by the number of uses, thus arriving at the frequency of failure but not the probability of failure. McCloy, Byrne and Johnson (2010) replicated these results in a series of new problems, clearer phrasings, and the use of a frequentist paradigm. Participants still made illogical and incorrect judgments, demonstrating the stability of Doyle's results.

To summarize the above review, people show a great deal of inaccuracy in many forms of probability judgment. This is especially true when they are attempting to estimate probabilities that involve more than one event. Kahneman and Tversky explained this using a heuristics model predicated on the idea that participants have preconceptions about their cognition and about social realities which lead them to make logically and mathematically inaccurate judgments. However, this is not the only explanation that has been presented for why people tend to do poorly when making probability judgments. In the next section, I will discuss the conjunction fallacy in detail as an example for the different explanations which can be given for poor probability judgment performance.

The conjunction fallacy

The conjunction fallacy was first described and studied by Tversky and Kahneman (1983). Probabilistic conjunctions refer to the combination of two events both happening, and the probability given for that co-occurrence. That is, given event A and event B, the conjunction is when both A and B occur. The probability of event A occurring can be called the marginal to the conjunction (this is also true for the other event, event B). If the probability of each marginal is estimated, then the probability of both events A and B occurring must necessarily be equal to or less than the probability of the least likely marginal occurring. This is because the overall probability of event A occurring also includes times when event B occurs and vice versa, but P (A and B) only comprises a subset of the marginal probabilities. When an individual estimates a conjunction at a higher probability than a marginal, that individual has committed a conjunction fallacy or conjunction error. For example, if someone estimated the probability that it would rain this week (marginal) against the probability that it would rain on Tuesday (conjunction), the probability of rain on Tuesday must be less or equal to the probability of rain during the week, since Tuesday is part of the week. Estimating the probability of rain on Tuesday as higher is irrational.

Tversky and Kahneman (1983), however, found that most participants did not follow this rule of logic when given a particular kind of problem. Participants in the study were given a series of problems which have become famous under the umbrella term of ‘the Linda problem.’ While there were other problems used in the same study, the Linda example has become the best known and most cited. Participants were given this text:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. (Tversky and Kahneman, 1983, p. 297)

Participants were then told to rank order a series of possible careers statements about Linda from most likely to least likely. There were eight possible options, but only three were actually important for the purpose of the study:

> Linda is active in the feminist movement. (F)
>
> Linda is a bank teller. (T)
>
> Linda is a bank teller and is active in the feminist movement. (T&F) (Tversky and Kahneman, 1983, p. 297)

85% of participants ranked F as most likely, but ranked T&F as more likely than T. This would appear to be a logical fallacy – as T&F is a subset of T, it cannot mathematically be more likely than T. Kahneman and Tversky (1983) also did several further experiments based on this conjunction fallacy, and also provided some possible explanations related to representativeness for why participants consistently make an error when given a probability conjunction. These findings and theories will be discussed in one of the later sections, but before continuing to those, I would like to discuss the flip side of the conjunction fallacy: the disjunction fallacy.

The disjunction fallacy

While the disjunction fallacy has several different definitions, the one I use in this paper wasdescribed by Bar-Hillel and Neter (1993) as the converse of the conjunction fallacy. A disjunction refers to an ‘either…or’ statement, which is true if either event A or event B occurs. A disjunction must, by definition, be greater or equal in probability to the most likely marginal, since a disjunction is true if either of those events occur. Both marginals are subsets of the disjunction event A or B. When an individual ranks event A as more probable than event A or B, or the same with event B, then that person has committed a disjunction fallacy or a disjunction error. For example, to return to our weather problem, someone might estimate the probability of rain this week (marginal 1) and the probability of wind this week (marginal 2). If they estimated the probability of wind or rain this week (disjunction) as lower than the probability of wind, they would be committing an error – the probability of wind or rain encompasses the probabilities of both disjunctions.

Despite this logical rule, Bar-Hillel and Neter (1993), similarly to Tversky and Kahneman (1983), found that most people did commit an error when given a certain type of problem that included disjunctions. Participants in their study were given a series of descriptions of different people, and were then asked to make probability judgments about that person. There is no one well-known example problem, as with the Linda example, but a sample problem given to participants involving 'Danielle' reads as follows:

> [Danielle is] sensitive and introspective. In high school she wrote poetry secretly.
> Did her military service as a teacher. Though beautiful, she has little social life, since she prefers to spend her time reading quietly at home rather than partying. What does she study? (p. 1122)

Participants were then asked to rank four areas of university study: Literature, Humanities, Physics, and Natural Science. In this example, Humanities and Natural Science represent disjunctions. Humanities is a superordinate category which includes Literature (event A) and everything else (event B). Natural Science is also a superordinate category which includes Physics (Event A) and everything else (event B). A participant who ranked Literature higher than Humanities – or Physics higher than Natural Science – would be committing a disjunction fallacy. Costello (2009a) has suggested that the disjunction fallacy can be considered in tandem with the conjunction fallacy. His research showed that participants who committed conjunction fallacies also tended to commit disjunction fallacies, and that problems which had high rates of conjunction errors across participants would also have high levels of disjunction errors when the problem was rephrased to fit the disjunction paradigm.

The research done by Bar-Hillel and Neter (1993) involves several manipulations, so there is no single disjunction fallacy rate at any point. However, across all conditions in their first experiment, participants committed a disjunction fallacy at least 50% of the time. The manipulations in this experiment and the possible explanations for different fallacy rates is discussed in one of the following sections, as I now move to the various attempts at explaining the conjunction and disjunction fallacies.

Theories of conjunction/disjunction

There have been several attempts at explaining these two, apparently related, probability fallacies. For the purposes of this paper, I have divided the explanations into four categories: social explanations, which use heuristics and stereotypes to explain the fallacy; problems with semantics, which argue that the Linda problem or the standard logical solution to the problem is somehow flawed and thus misconstrues an actually reasonable response by participants; numeracy and cognitive error explanations, which propose that individuals have a general difficulty with probability and mathematics which is not limited to the social context of the conjunction fallacy; and the integration explanation, a subset of the numeracy explanations, which suggests that individuals have a more specific problem with integrating two probabilities together in order to create accurate assessments for conjunctions or disjunctions.

Social explanations

Tversky and Kahneman (1983) explained the original conjunction error findings using the representativeness heuristic, which had already been used in explaining the base rate fallacy and the anomalies in whether something seems ‘random’ or not (Kahneman and Tversky, 1973; Tversky and Kahneman, 1974). The representativeness heuristic refers to how much a sample fact seems to stereotypically fit the person the fact seems to be describing. In the case of the Linda problem, Linda has been described as a socially active and humanities-oriented student when she was in college. Given these descriptors, it seems more stereotypical that Linda would be a feminist than that she would be a bank teller, so that is ranked higher. That judgment, in and of itself, is not logically contradictory. However, representativeness does not play well with conjunctions. In this case, bank teller is not at all representative of Linda’s personality – there is nothing in her history that makes her seem likely to become a bank teller. However, a feminist bank teller does contain traits that are representative of Linda’s description. Even though the ‘bank teller’ marginal appears to be socially implausible, its conjunction with feminist paints a more compelling picture, even though it is logically inconsistent to rank the conjunction over one of the marginals.

As seen above, the results from Tversky and Kahneman (1983) appear to fit this representational heuristic: 88% of all participants made a conjunction error on the single problem they were tested on. While this rate is collapsed across two sample problems and three subject populations with varying levels of mathematical knowledge, there is little deviation from this rate between conditions. Tversky and Kahneman performed other manipulations, looking for a case in which participants rated conjunctions logically lower than their components, but the fallacy remained stable. The conjunction error was not significantly diminished by removing the foils (the other options for Linda that were not being coded). Even when participants were explicitly presented with two competing arguments (one using the logical rule, the other using the representational rule), 65% of participants still relied on the representational rule. Given the stability of the conjunction fallacy, and the participants' apparent endorsement of the representational rule, it would appear the Tversky and Kahneman argument is a good one. Representativeness is certainly prevalent in which marginal characteristic is predicted as more or less likely – given the fact that participants are only given information about the character, not about base rates in society or any other information, the 'fit' of one option to the described character is the only way in which participants can differentiate between the options. If there was no element of representativeness, thus, we would expect Linda to be rated as a feminist just as often as she is rated as a bank teller. The fact that Linda is overwhelmingly perceived as a feminist rather than a bank teller goes some way to confirming Tversky and Kahneman's theory.

After the first study by Tversky and Kahneman (1983), several researchers used the representativeness model to further explore the conjunction/disjunction fallacies. Bar-Hillel and Neter (1993) used socially-relevant cues and descriptions to test the disjunction fallacy, in a similar paradigm to the Tversky and Kahneman (1983) study. Bar-Hillel and Neter suggested that high fallacy rates among participants were due to participants choosing the options that appeared most representative of the character they had been given. In a more recent study, Barch, Schultz, Chechile and Sommers (2012) confirmed the conjunction fallacy in a socially-relevant task, but found almost no errors in a toy marble task that had no social relevance, supporting the idea that the conjunction fallacy is predicated on social heuristics and stereotypes. Other research has shown that autistic adolescents appear to be less susceptible to the conjunction fallacy than neurotypical adolescents, presumably because that group is less likely to process the social cuing that is hypothesized to create representativeness (Morsanyi, Handley and Evans, 2010). The same study found that conjunction errors are less frequent in cases where neither marginal trait is particularly representative of the assigned character, as opposed to the traditional Linda problem where one marginal trait is very representative and the other is not. Finally, Fisk and Pidgeon (1997) offer a twist on the representativeness hypothesis, suggesting that participants begin from the unlikely marginal, assigning a low weight to it depending on how ‘surprising’ and non-representative it seems, and then devalue the conjunction based on this non-representativeness. However, since participants initially give the conjunction a higher weight because it contains a representative element, they still commit a conjunction error by rating the conjunction higher than the unlikely marginal. Testing this hypothesis found that participants do indeed assign a higher surprise factor to the unlikely marginal as well as to the conjunction (Fisk and Pidgeon 1997; Fisk and Pidgeon 1998).

To summarize the argument across these studies, these researchers have hypothesized that the conjunction/disjunction fallacy is related to representativeness and social processing. People commit the conjunction error because they are looking for what seems to *fit* an individual or a situation the best, rather than trying to create a logically consistent set of probabilities. If this hypothesis is true, then the conjunction/disjunction error will not be present in situations which do not have an associated heuristic or stereotype which would influence the rating of marginals and the conjunction as more or less representative of the person or situation being judged.

Semantic explanations

However, there are some problems with the social explanations for the conjunction/disjunction fallacy. A common objection from other researchers has been that there are inherent problems with the “Linda” paradigm which lead the participant to provide a logically incorrect answer. These objections pertain, by and large, to three factors: the use of ranking to differentiate between options, the use of the word ‘probability,’ and the clarity of the conjunction option.

The use of ranking in the Tversky and Kahneman (1983) study casts the implications of their results into doubt. As Costello (2009b) points out, if P (B) and P (A and B) are extremely close together or even equal to each other, only a small amount of noise or random error in the probability judgment is required for the participant to rank P (A and B) higher than P (B). If the probability of the conjunction is considered equal to the probability of one of the marginals, the participant might be logically consistent in ranking the conjunction probability ‘higher’ than the marginal probability, given that there is no ‘equal to’ option provided in the traditional Linda paradigm. This problem is also present in the Bar-Hillel and Neter (1993) study on the disjunction fallacy, which also used ranking when asking for participant responses. A subset element might be ranked higher than the disjunction because a participant thought they were equal in probability.

This distinction may seem to be minor, but Morier and Borgida (1984) showed that it actually matters quite a bit. They compare the ranking paradigm from Tversky and Kahneman’s classic study to a new rating system where participants were asked to provide a numerical probability estimate. This allowed the researchers to distinguish between logically-consistent ties and actual conjunction errors. Using the original Linda problem, including the foil answers, Morier and Borgida found that the conjunction error rate was reduced from 95.2% in the rank order condition to 80% in the probability estimate condition. While the error rate is still quite high in both conditions, changing the rating procedure did significantly change the number of errors. This suggests that some degree of the conjunction fallacy can be traced to the formulation of the task.

The use of the word ‘probability’ within the Linda problem is also suspect. Participants in the original Tversky and Kahneman (1983) study were told to rank the possible outcomes of the Linda problem according to their probability, but the word probability was not defined. Hertwig and Gigerenzer (1999) showed that probability is a polysemous word – it has many meanings, and some of the most common ones are not mathematical. Rather, Hertwig and Gigerenzer found that most participants interpreted probability in this case as having to do with a sort of ‘goodness of fit,’ thus priming the representativeness heuristic. Hertwig and Gigerenzer suggest that using a frequency model would produce a clearer picture of people's actual ability at make logically consistent mathematical estimates. Instead of being asked for the difficult conception of probability, participants would be asked for the frequency of a woman fitting Linda's description pursuing bank telling, feminism, or both. Unfortunately, this claim has received mixed results. A preliminary experiment by Hertwig and Gigerenzer (1999) with relatively few participants showed that using a frequency model for the Linda problem sharply reduced conjunction errors, from a low of 82% in the probability conditions to a high of 20% in the various frequency conditions. But later research using weather conjunctions has only shown a reliable but much smaller decrease in both conjunction and disjunction errors between the traditional probability model and the frequency model (Costello, 2009a). Sides, Osherson, Bonini, and Viale (2002), meanwhile, found no reliable difference between asking for 'probability' and asking for participants to make a bet without mentioning probability at all. While the use of 'probability' as a descriptor may explain some of the conjunction/disjunction fallacy, there seems to be no consensus on how much of an effect it actually has on participants.

Finally, the clarity of the conjunction statement itself is also in question. The argument is that presenting subjects with three options – event A, event B, and event A and B – implies the 'A' is equivalent to 'A but not also B,' and 'B' is equivalent to 'B but not also A.' Participants may thus believe that they are acting logically in ranking the conjunction over the marginals. It seems more likely that Linda would be a bank teller and a feminist, rather than a bank teller that is not in any way involved with the feminist movement. However, this idea has not received substantial support. Bonini, Tentori and

Osherson (2004) performed a betting variation of the conjunction task which explicitly informed participants that a conjunction referred to both events happening, and found no significant difference between this and earlier studies in the paper which had made no clarification. A later variation within the same study which provided participants with three options – 'A,' 'A and B,' and 'A and not B' – also saw a high level of errors. Sides et al. (2002) used other forms of the conjunction such as 'A, after which B' and told participants the 'A' meant any form of 'A.' This also did not reduce conjunction errors. The argument that participants do not really understand the meaning of the marginals or conjunctions does not appear to have been borne out.

In summary, the argument for semantic incoherence suggests that problems with the phrasing in the Linda problem produce a flawed result with an artificially high level of errors. It is suggested that if the problem was phrased in a more easily understood manner, there would be significantly fewer errors, and perhaps that the conjunction fallacy would cease to be a problem. However, attempts to prove this by rephrasing and refining the conjunction problems have met with little success. While it can be seen that some problems (particularly ranking instead of using probability estimates) do create more errors, the entire conjunction/disjunction fallacy cannot be explained by problems with the task itself, and conjunction errors persist even in carefully phrased and tweaked experiments.

Numeracy and cognitive errors

In response to the conjunction/disjunction fallacies and the other errors in probability judgment discussed above, some researchers have conjectured that people have a fundamental difficulty with making probability judgments. The theory is that the errors made in conjunction and disjunction problems are generalizable to more systematic problems in probability judgment overall.

Valerie Reyna has been at the forefront of work on numeracy, or the ability of people to make accurate and logical judgments based on numbers. Numeracy, like literacy, is something which people can be trained in in order to help them comprehend numbers and mathematics better. People with low numeracy scores tend to make a consistent and high amount of mistakes when dealing with probability, while people with high numeracy tend to make relatively few mistakes (Reyna et al., 2009). If applied to the conjunction fallacy, it can be suggested that the participants who are making conjunction errors are simply having difficulty processing the numbers involved in making the probability judgments. While the original study on the conjunction fallacy by Tversky and Kahneman (1983) found that expert participants in the medical field with presumably high numeracy still made conjunction errors, the numeracy hypothesis is worth considering in light of the procedural problems with that original study, as discussed in the last section.

Other researchers have emphasized overall problems with cognition that negatively impact everyone's judgment of probability. Gavanski and Roskos-Ewoldsen (1991) demonstrated similar levels of conjunction errors in both socially-relevant and socially-non-relevant conjunction problems. They found that participants in their conjunction study tended to use incorrect rules to integrate the marginals into a conjunction. Thus, while participants may or may not have been judging the marginal probabilities correctly or using representativeness to judge some marginal probabilities, the major factor in conjunction errors appeared to be mistakes in the complex probability calculation, which could be a problem even in tasks that had nothing to do with social situations. A similar finding is that people perform poorly when calculating cumulative probabilities, or the probability of an event occurring over a long span of time. Participants incorrectly produce judgments that conform to frequency judgments instead of the logical probability judgments they were asked for (McCloy, Byrne and Johnson, 2010; Doyle 1997).

The theory that people are commonly making some sort of basic mistake in probability calculations themselves is supported by the reaction of participants when they are corrected or when they receive feedback in some other manner. Bar-Hillel and Neter (1993) found that participants in their disjunction study who committed an error would later recognize that the superordinate categories included the more specific categories. When participants were told the results of their bet in the third experiment – where a betting paradigm with real money was used to test the disjunction effect – participants saw that they had made a mistake by betting on the specific category, rather than the superordinate category. Indeed, when groups work on conjunction problems together, the existence of a single person in the group who understands conjunction problems can greatly reduce the rate of errors. Such groups do not act on a false consensus, but use the knowledge of the minority expert (O'Leary, 2011).

In summary, the argument for a general error in the conjunction/disjunction fallacy suggests that difficulties in solving these problems are part of a general human difficulty with the conceptualization of probability. Participants tend to default to frequency judgments or completely different judgment rules when asked about probability, and the level of expertise with numbers and calculations (numeracy) appears to matter a great deal. However, when participants are corrected, they can easily recognize what mistake they have made and adjust their judgment accordingly. If this theory is true, we can hypothesize that people who make conjunction/disjunction errors will be consistent in making errors when not corrected, and will also make errors in other probability judgments. Participants should also have trouble with probability in any kind of task, not just in socially-relevant problems.

Probability integration

A theory that is related to the numeracy explanation is that people show a specific difficulty with probability. In this theory, errors in the conjunction/disjunction fallacy do not come from mistaken heuristics or even widespread errors in probability, but from a particular problem with integrating marginal or specific probabilities into a conjunction or disjunction.

Evidence for this theory has been shown in a range of research. Nilsson (2008) showed that participants committed more conjunction errors when they were given more possible conjunctions in a single problem and thus had to perform more conjunction calculations. As discussed in the last section, Gavanski and Roskos-Ewoldsen (1991) found that in their conjunction study, participants tended to make errors during the conjunction stage, rather than when assessing marginal probabilities, and participants did not assess conjunction probabilities separately from the marginals. This suggests that the conjunction errors are made while combining the probabilities.

There are several suggested explanations for why participants are combining probabilities incorrectly. Some participants appear to combine probabilities based on frequency of occurrence, which can result in incorrect calculations or correct frequency calculations which simply fail to answer the probability-based problem (Doyle, 1997). This suggests that participants simply do not know the correct way to combine the marginal probabilities. The difference in probability between marginals may also contribute to participants' trouble. Conjunction errors are more common when one marginal has very high probability and the other marginal has very low probability (Gavanski and Roskos-Ewoldsen, 1991). If participants are simply using the wrong mathematical rule to perform conjunctions, then a greater difference between the marginal probabilities could exacerbate the error and make it more readily apparent.

Another suggestion is that participants are attempting to use Bayesian updating in order to answer conjunction problems (Crupi, Fitelson and Tentori, 2008). In the Linda problem, a participant is led to assume that the likelihood of Linda being a feminist is very high, while the probability of her being a bank teller is very low. But receiving the information that she may be a feminist in addition to being a bank teller, participants update their expectations about the less likely bank teller hypothesis to include the positive evidence from feminism, and thus rate the conjunction higher. While this is the incorrect logic for the specific Linda problem, it is the correct logic for trying to deduce the identity of a person or object based on cues or traits, as seen, for instance, in a game of 'Guess Who.' If we are trying to find a suspect and we receive new information that he has shaven his mustache, we must update our probabilities of whether he is the man with a full beard or if he is the man who is newly clean-shaven. This Bayesian logic was also seen in studies done by Wolford, Taylor and Beck (1990). In these studies, some conjunction problems could be correctly answered using Bayesian updating – participants had to estimate the probabilities of one of three class registration cards belonging to their friend. Participants generally answered these conjunction problems correctly, but also used the same logic to incorrectly answer traditional conjunction fallacy problems which are not amenable to Bayesian updating. The Associative Learning paradigm, created by Cobos, Almaraz and Garcia-Madruga (2003), also uses Bayesian updating to explain the conjunction fallacy. The researchers hypothesize that participants gave different predictive weights to the various cues present in the Linda problem, and that the order of the cues and the options given then affect how the probability judgments are updated. The experiments in their paper bear out the suggestion that the predictive strength of given cues and the order in which participants are asked to consider a conjunction can make a difference in the number of conjunction fallacies.

Finally, it has also been suggested that participants are confused by the apparent causal connection in some conjunction problems, particularly those to do with real world predictions. When Event A makes Event B more likely, it is tempting to think that they are more likely in conjunction then separately: cold rain seems more likely than rain without cold. Participants might thus be using causal reasoning when making the conjunction error, in addition to the false ‘either or’ problem which leads participants to believe that Event A excludes cases where A and B occur together. Crisp and Feeney (2009) found that conjunction fallacies were more common when the component events were causally connected than when they were unrelated.

Regardless of where the difficulty with integrating probabilities originates from, this explanation does have potential as the cause of the conjunction/disjunction error. If there is something specifically wrong with people’s ability to combine probabilities, we would expect to see less accuracy when people judge conjunction or combined probabilities than when they judge marginal or specific probabilities. This will be even more evident when integration is made more difficult or demanding.

Conclusions

As the above literature review shows, work in probability and specifically in the conjunction/disjunction fallacy has been wide-ranging, but still has significant holes. Four competing explanations for conjunction/disjunction errors have emerged, each with their own prerequisites and predictions.

The social explanations argue that conjunction/disjunction errors have to do with socially-relevant heuristics and stereotypes. In this theory, participants rate conjunctions higher than marginals because the conjunctions appear to 'fit' the example person better. This theory predicts that conjunction errors only occur in socially-relevant contexts. In a socially-relevant problem, participants should be expected to make a large number of errors. In a problem which does not active social heuristics, however, there should be few or no errors.

The semantics explanation proposes that conjunction/disjunction errors arise from poor formulation of the task leading to participant misunderstanding. This theory suggests that the number of errors will depend on the wording of the task. In a task which uses words such as 'probability,' for instance, there should be more errors than in a task which asks for frequency judgments.

The numeracy explanation argues that people have a general difficulty with probability and numbers which is not limited to the conjunction/disjunction fallacy. If this is true, most participants should be inaccurate all around in judging probability, including when judging marginal probabilities as well as when judging conjunctions.

Finally, the more specific probability integration theory proposes that people commit conjunction/disjunction errors because they have difficulty combining probabilities into a conjunction or disjunction. This would predict that participants will be more accurate when making judgments about probabilities that require no integration than when making judgments about integrated probabilities. In addition, it suggests that participants will make more errors when performing a task that requires more integrative work.

In the following pages, I outline a study which can discriminate between these four explanations and produces conditions where variations in conjunction/disjunction errors and general accuracy in probability judgment can test the above predictions for the conjunction/disjunction fallacy.

III. Experiment 1

In the above literature review, I identified four different theories regarding the conjunction fallacy, each with their own predictions about how error rates and general probability accuracy will vary depending on what form of conjunction/disjunction task the participants are given. In experiment 1, I operationalized these variations in order to create a task which can discriminate between the four hypotheses. Experiment 1 used a toy marble task with no known associated social heuristics. The marble task did, however, include several manipulations for level of integration necessary to complete the problems. Finally, a variety of different probability estimations allowed me to make a distinction between probability judgment accuracy and conjunction errors. Given these manipulations, the social explanations, the numeracy explanation, and the probability integration explanations all had different predictions for the outcome of experiment 1. In this section, I will briefly explain the reasoning behind the formulation of the probability judgment task and manipulations I used before going into the study itself.

Experiment design

The marble task

The same marble task was used in both experiments reviewed in this paper. The marble task had been previously developed by Barch, Schultz, Chechile and Sommers (2012) for other research on the conjunction fallacy in this lab. By using a simple task without associated stereotypes, their research and this experiment can differentiate between the social explanation for the conjunction/disjunction fallacy, which would predict no errors, and the other theories which do not rely on heuristics and stereotypes to explain the fallacy.

In the standard marble task, participants were presented with the image of forty marbles, hovering over a gray can. Each marble had one of two different colors, and one of two different patterns: red and blue, and striped or solid. Thus, there are four marginal categories of marbles (red, blue, striped and solid), and four marble conjunctions (red solid, blue solid, red striped and blue striped). After a brief delay, the marbles dropped out of sight, into the can. After this presentation, participants were asked two questions about the marbles, in the format of “what is the probability of picking a [target marginal or conjunction] marble from the can?” After participants answered the two questions, they were presented with a new field of marbles.

Comparing the participants’ subjective probability judgments to the objective probability of randomly picking the target marble provides a measure of the participants’ basic accuracy. But pairing the two questions in a certain way also provides a measure of conjunction errors without asking every relevant question and possibly cuing participants as to the exact purpose of the experiment. By asking one marginal question and one conjunction question, we can infer the rest of the relevant probabilities and detect a conjunction error.

	Blue	Red	
Striped	A (Blue Striped)	B (Red Striped)	E (All Striped)
Solid	C (Blue Solid)	D (Red Solid)	F (All Solid)
	G (All Blue)	H (All Red)	

Fig. 1
Question Matrix

In the two-question paradigm, participants are asked one conjunction question (A, B, C, or D) and one marginal question (E, F, G, or H). (See Fig. 1 for a clear guide.) For instance, a participant might be asked about the probabilities associated with A and F, or blue striped and solid. When the participant estimates the probability of F, the probability of E is implied by the marble dichotomy. Since any marble that is not striped must be solid, the probability of E=1-F. In the traditional question paradigm seen in the Linda problem (Tversky and Kahneman, 1983), participants are asked every question relevant to the expected conjunction fallacy. Thus, the given probabilities to A and E would be compared, and if A were more than E, a conjunction error would be committed. There cannot be a higher probability of getting a blue striped marble than a striped marble. The two-question paradigm is more indirect. By using the implied probabilities, we can see that if A is more than 1-F, there is a conjunction error. In the marble paradigm, this calculation also measures disjunction errors: if we take ‘striped’ as a superordinate category, then estimating the probability of the more specific ‘blue striped’ above the implied probability of ‘striped’ would be a disjunction error. While this implied measurement is potentially more imprecise than direct comparisons, it allows me to minimize the number of questions and time taken on each trial while not alerting participants to the hypothesis. This method has been previously used to successfully detect conjunction errors. Barch et al. (2012) used the two-question paradigm to detect conjunction errors in a complex, socially-relevant condition using crime statistics and different genders and ethnicities. In another condition using the simple, non-socially-relevant marble task, however, the researchers did not detect any conjunction errors. This method, thus, has already been shown to be able to detect errors and to be sensitive to differences between conditions.

Manipulations

In experiment 1, I manipulated the basic marble task to make some conditions more difficult than others in terms of probability integration. In the probability integration explanation for the conjunction/disjunction fallacy, participants should be less accurate and make more errors when they experience conditions requiring more integrative work. The other theories, however, do not have specific predictions for the integration conditions. I manipulated delay between presentation and test, and how the marbles were presented, in a 2x2 factorial design. In this section, I explain the two manipulations and how they affect probability integration.

Test-delay

Having to recall and review information can make integrating that information more difficult. Crisp and Feeney (2009) found that participants operating under a heavy memory load had more difficulty making accurate conjunctions than participants with no added memory load. One possible explanation for this effect is that participants are losing their exact, verbatim knowledge of the problem and instead relying on a more intuitive process. Fuzzy Trace Theory, for example, argues that people often rely on 'fuzzy' gist memories of things in order to make judgments about them (Wolfe and Reyna, 2010; Miller and Bjorklund, 1998). These gist memories persist after the exact memories have disappeared. If participants lose their exact memories of the marble field after some delay, they will have to integrate the gist memories of the marbles together to reach some decision, potentially making their judgments less accurate.

In this experiment, I manipulated memory load by a test-delay manipulation. In the delay condition, participants were shown the field of marbles, and then had to perform a filler task for two minutes before being asked the questions about marble probability. In the filler task, participants completed words of either four or five letters: participants were given a word with two or three letters missing, and told to fill in the blank. Participants were allowed to skip words if they could not think of an answer, but they were also encouraged to fill out every word to the best of their ability. At the end of the filler task, participants were reminded to only consider the last set of marbles they saw when answering the probability questions. Participants completed 20 of these delay tasks.

In the control condition, there was no delay or filler task between presentation and test, and participants were immediately directed to give their probability estimates. Participants completed 20 of these no-delay tasks.

Presentation

Presentation of information is another factor in how much integration is required of participants. More complex conjunction problems with more information require more probability integration (Nilsson, 2008). Tversky and Kahneman (1983) used an extremely complex problem for the Linda problem – participants were given a brief biography of a novel person, and required to make judgments about her current life based on her past history. According to the representativeness argument, the high fallacy rate associated with the traditional Linda problem is due to the heuristics activated by Linda's life story. But according to the probability integration theory, the sheer number of cues about Linda's life contributed to the difficulty of bringing all of this information together in order to produce a single probability estimate. Problems with fewer pieces of disparate information still show a conjunction/disjunction fallacy, but to a much smaller degree (Costello, 2009a). The marble problem is inherently less complicated than the Linda problem, since all the information needed is present on the screen. However, if presentation is made more complex, the integration hypothesis would expect that participants would make more errors than when presentation is simpler.

In this experiment, I manipulated presentation complexity by changing how marbles were shown. In the segment condition, the marbles were presented in segmented groups, rather than as a complete field. Six or seven marbles were presented for one second, before falling into the bucket. The first group was followed by a second, and then a third, and so on for six groups, totaling forty marbles. Participants were reminded to consider all of the marbles that fell into the bucket when making probability judgments about the presentation field. Participants completed 20 of these segmented tasks.

In the control condition, all forty marbles were presented at once, before falling into the bucket. The time for both the sectional presentation and the whole presentation was held constant between conditions, so the whole marble presentation is shown for six seconds before falling into the bucket. Participants then proceeded to the probability questions. Participants completed 20 of these control tasks.

Summary of manipulations

In summary, there are two manipulations of test-delay and presentation complexity. These manipulations were crossed to create four conditions.

In condition 1 (segmented and delay), participants were presented with a series of 6 marble groups, with six or seven marbles each for a second at a time. After all six groups had fallen into the bucket, participants performed a word-fill delay task for two minutes before answering two probability questions about the marbles. Participants performed this task ten times.

In condition 2 (segmented and no delay), participants were presented with a series of six marble groups, as before, but performed no delay task before answering the probability questions. Participants performed this task ten times.

In condition 3 (whole and delay), participants were presented with a field of forty marbles for six seconds. After the marbles fell into the bucket, participants performed the delay task for two minutes before continuing to the probability questions. Participants performed this task ten times.

In condition 4 (whole and no delay), participants were presented with the forty marbles as before, but performed no delay task before answering the probability questions. Participants performed this task ten times.

In total, there were forty trials, ten for each condition. Participants did the 20 whole presentation trials first (conditions 2 and 4), randomly mixed between delay and no-delay. Participants then completed the 20 sectional presentation trials (conditions 1 and 3), again randomly mixed between delay and no-delay.

Hypothesis and predictions

I devised this experiment primarily to test the probability integration theory. Based on that theory, I hypothesized that participants would commit more errors in the high-integration conditions: delay and segmented presentation. I also hypothesized that participants would be less accurate when judging marginals than when judging conjunctions, as in this task all of the marbles presented on the screen are conjunctions and participants must integrate two conjunction probabilities together to reach a marginal estimate. For example, red/striped and red/solid marbles must be integrated together to create an estimate for red marbles. Following the probability integration theory of the conjunction/disjunction fallacy, participants should be less accurate when more integration is necessary. If participants perform equally as well or better in the delay and segmented conditions than in the control conditions, this would discredit the probability integration hypothesis. Similarly, if participants are equally accurate when judging conjunctions and marginals, this would disagree with the hypothesis.

This experiment also serves as a test of the social explanations for the conjunction/disjunction error. In this experiment, participants were presented with a toy marble problem, which should have activated no heuristics or social stereotypes. The social explanation, thus, would expect no conjunction/disjunction errors. If participants do commit conjunction/disjunction errors in the marble task, this would discredit the social explanations.

While the experiment contains no rigorous test of the semantic explanations for the conjunction/disjunction fallacy, it does use the word ‘probability,’ which has been suggested to increase conjunction/disjunction errors. The semantic explanation has already been mostly proved to be an inadequate explanation for the entire conjunction/disjunction fallacy (Bonini, Tentori and Osherson, 2004; Sides et al., 2002). This experiment can extend that research by seeing if there are conjunction/disjunction errors in this case, where they are tentatively predicted by the semantic explanations but not by the social explanations.

Finally, this experiment also can test the numeracy explanation for the conjunction fallacy. If participants are having difficulty with numbers and calculations, I would expect accuracy to be low on both marginals and conjunctions, and participants should make conjunction/disjunction errors in all conditions. The predictions, thus, are different for each theory.

Method

Participants. The participants were twenty-one undergraduates at Tufts University. Individuals participated in exchange for fulfillment of course criteria in introductory psychology classes. Participants did not receive any other compensation for taking part in the study.

Materials. Images of the marbles and screenshots from the experiment can be found in Appendix A.

Procedure. Participants completed the forty marble tasks described above on a computer. The experiment took about fifty minutes, with some participants taking more or less time to complete the experiment.

Results and Discussion

Data was coded for several criteria: the two manipulations of presentation and delay, the question types of marginals and conjunctions, and the number of conjunction/disjunction errors. Difference scores for each probability judgment were found by subtracting the subjective judgment from the objective probability of getting a certain type of marble, and then taking the absolute value of that. All of the tests below used these difference scores.

Probability accuracy

	Whole Presentation	Sections Presentation
Delay	10	10
No Delay	10	10

Fig. 2
Median difference scores for the crossed conditions

When comparing participants' accuracy in judging probability, none of the condition medians were significantly different from each other (see Fig 2). The means of the conditions (with standard deviations in parentheses) were as follows: Whole x Delay, 14.39 (13.02); Whole x No Delay, 11.25 (9.98); Segmented x Delay, 13.08 (11.8); and Segmented x No Delay, 10.52 (10.34). A factorial ANOVA found a significant main effect for the delay manipulation, *F*(1, 1,656)[2] = 28.27, p < .0001. The presentation manipulation was not significant, *F*(1,

[2] The degrees of freedom may seem somewhat high. This is because there were an extremely high number of observations that were used for the ANOVA. 21 participants completed 40 trials each, in which they were asked to make 2 judgments per task. This resulted in 1,680 observations, with 1,659 for the total

1,656) = 3.60, p = .058, nor was the presentation x delay interaction, *F*(1, 1,656) = .29, p = .5879. While there is a discrepancy between the medians and the mean testing, this is probably due to noise when participants estimate particularly low or particularly high probabilities.

Question type

Marginals	Conjunctions
10	10

Fig. 5.
Medians of the marginal and conjunction difference scores.

Difference scores were also used to compare participant accuracy when judging marginal probability to their accuracy judging conjunction probability. Originally four questions were coded separately for each participant – two where participants were asked two marginal questions together, and two where participants were asked two conjunction questions together. However, the mean and median scores for these marginal and conjunction questions were not apparently different from the other marginals and conjunctions, so those scores were collapsed into the two larger measures. The median difference scores for all marginals and all conjunctions, as seen in fig. 5, were exactly the same. The means were marginals (M = 12.78, SD = 12.39) and conjunctions (M = 11.75, SD = 10.6). There was no significant difference between these two means, *t*(439) = 1.28.

Conjunction/disjunction errors

Out of 756 trials that had a possibility of a conjunction/disjunction fallacy, 16 contained conjunction/disjunction errors, or an across-participants rate of about 2%. Out of the 16 errors recorded, 1 participant committed 3 errors, 4 participants committed 2 errors, 5 participants committed 1 error, and 11 participants committed no errors at all. 7

degrees of freedom. After subtracting for subjects (-20) and the manipulations (-1 for delay, -1 for presentation, and -1 for the interaction), this leaves us with 1,656 degrees of freedom.

errors were committed in the segmented presentation condition, and 9 were committed in the whole presentation condition – a relatively even split. However, only 4 of the errors were committed in the no-delay condition, with 12 errors committed in the delay condition. A matched-pairs randomization test on the number of errors participants committed in delay vs no-delay conditions showed the result to be non-significant (P = .09).

Experiment 1 discussion

These results are largely consistent with the social explanations for the conjunction effect: in a toy problem with no social element, participants made almost no conjunction errors and were generally within 10% of the objective probabilities. This is contrary to my integration-based hypothesis: participants did not generally commit more errors in the high-integration conditions, nor were they less accurate when judging marginals than when judging conjunctions. The results are also inconsistent with the numeracy hypothesis, as participants were highly accurate when judging probability. However, there may have been some negative effect from test delay on probability accuracy – when comparing the means, participants were somewhat less accurate in the delay condition, but this was not the case when comparing the medians. In experiment 2, I repeat the delay manipulation to attempt to clarify the impact of delay on probability judgment.

IV. Experiment 2

Experiment 2 was designed to confirm the basic results from experiment 1 while adding a new factor: the generation effect. In this section, I define the generation effect before going on to discuss the experiment design.

Generation literature review

The Generation Effect was first described by Slamecka and Graf (1978). Their series of experiments showed that when participants actively created a response instead of passively reading information, they were more likely to remember information later on. Participants in the passive condition simply read word pairs such as 'rapid-fast.' Participants in the generation condition created their response by completing word pairs such as 'rapid-f___.' Since the rule in this experiment was that each pair had to be a synonym, active participants were constrained in their response and still generated the same word as the passive participants. However, recall and recognition tests showed that participants had a much greater memory for the generated response than for the passive response.

The generation effect has since been replicated and expanded in many successive studies. While the generation effect was first demonstrated with closely-related word pairs, it also serves as a powerful memory boost in many other areas. The generation effect has been demonstrated with incongruous word associations (Soraci et al, 1994; Chechile and Soraci, 1999), as well as wordlike and unwordlike nonwords (Johns and Swanson, 1988). More importantly for the following experiment, the generation effect has also been demonstrated with numbers and problem solving, where participants were more likely to remember the solutions to math problems if they had generated the answer themselves, rather than passively reading the answer (Gardiner and Rowley, 1984). Finally, the generation effect has also been shown to enhance the memory of visual images (Peynircioglu, 1989; Kinjo and Snodgrass, 2000). Thus, the generation effect is highly applicable to the marble task, which features both visual data (the marble presentation) and numerical judgments (the probability questions).

The research on the generation effect has reached some points of consensus. The generation effect results in increased memory for the generated response created by the participant, compared to the memory for the same response in the passive reading conditions (Slamecka and Graf, 1978; Gardiner, 1988). The memory for the cue or stem, however, is not increased (Slamecka and Graf, 1978).

Overall, the generation component is very easily added to the current research paradigm. In experiment 2, I added a new factor to the marble task to test the impact of the generation effect on probability judgment.

Manipulations

Experiment 2 used the same marble task as experiment 1 in a 2x2 factorial design, but without the presentation manipulation. I retained the test delay manipulation, but added a generation manipulation, as described below.

Generation manipulation

There were two conditions in the generation manipulation: the passive reading condition, and the generation condition. In the passive reading condition, participants were shown the marble presentation of forty marbles for six seconds. After the marbles fell into the bucket, participants read a factual sentence about the relative proportion of marbles. These sentences always took the form of “There were [more or fewer] [marginal] marbles than [opposite marginal] marbles. An example is “There were more red marbles than blue marbles.” Participants then continued to the probability questions. Participants completed 20 of these passive trials. In half of the trials, the marginals used were ‘red or blue,’ and in the other half they were ‘striped or solid.’ In half of trials, there were more of one marble than the other, and in the other half there were fewer of one marble than the other.

In the generation condition, participants were shown the marble presentation as usual, but, after the marbles fell into the bucket, participants were given an incomplete sentence about the marbles. The sentences followed the same format as the passive read sentences, but the word ‘more’ or ‘fewer’ was missing. For example, “There were ____ red marbles than blue marbles.” Participants were told to complete the sentence with either ‘more’ or ‘fewer.’ After they typed in an answer, they then proceeded to the rest of the task. Participants complete 20 of these generation trials. In half of the trials, the marginals used were ‘red or blue,’ and in the other half they were ‘striped or solid.’ Also, in half of trials, the correct answer was ‘more’ and in the other half it was ‘fewer.’ This meant that participants actually had to think about the answer – they could not simply type in ‘more’ for every generation statement and still be correct.

Test-delay

The test-delay manipulation was the same as in experiment 1. In half of the 40 trials, participants saw the presentation of forty marbles for six seconds, and then immediately answered probability questions pertaining to the presentation. In the other 20 trials, participants saw the marble field, and then performed the word-completion task for two minutes before answering the probability questions.

Summary of manipulations

In summary, there were two manipulations of generation and test-delay. These manipulations were then crossed to make four conditions.

In condition 1 (generation and no delay), participants were presented with forty marbles for six seconds, before they fell into a bucket. Participants were then asked to complete a generation sentence about the marbles, before continuing on to immediately answer two questions about probability. Participants performed this task ten times.

In condition 2 (passive and no delay), participants saw the marble field, but were then presented with a passive read statement about the marbles. After viewing the statement, participants continued on to immediately answer the probability questions. Participants performed this task ten times.

In condition 3 (generation and delay), participants saw the marble field, and were then asked to fill out a generation sentence. They then spent two minutes performing the word completion task, before going on to answer the probability questions. Participants performed this task ten times.

In condition 4 (passive and delay), participants saw the marble field, and were then presented with a passive read statement. They then spent two minutes on the word completion task, before going on to answer the probability questions. Participants performed this task ten times.

In total, there were forty trials, with ten from each condition. All four of these conditions were mixed randomly together, meaning that each participant saw the forty trials in a unique order. Participants did not know at any given time which condition the next trial would be from.

Hypotheses and predictions

Given the extensive research already done on the generation effect, I hypothesized that participants would make more accurate judgments about probability in the generation condition. They should also be less likely to commit a conjunction/disjunction error, although, given the low rate of conjunction/disjunction errors in experiment 1, there may be a ceiling effect where performance cannot improve any more than the high level participants showed in experiment 1. Finally, I hypothesized that participants would be less accurate in the delay condition than in the no delay condition, resolving the mixed empirical data from experiment 1.

Method

Participants. The participants were twenty-one[3] undergraduates at Tufts University. Individuals participated in exchange for fulfillment of course criteria in introductory psychology classes. Participants did not receive any other compensation for taking part in the study.

Materials. Screenshots of the generation manipulation can be found in Appendix B. The marble images, whole marble presentation, and wordfill task screenshots can be found in Appendix A.

Procedure. Participants completed the forty marble tasks described above on a computer. The experiment took about fifty minutes, with some participants taking more or less time to complete the experiment.

[3] Twenty-two students actually participated in the experiment. Unfortunately, due to technical difficulties, we lost data for a few trials after one participant completed the experiment. Their data is not used in the following analyses, reducing the *n* for participants to twenty-one, or 840 trials.

Results and Discussion

Once again, I used difference scores and coded participants' data along several criteria: the two manipulations of generation and delay, the question types of marginals and conjunctions, and the number of conjunction errors.

Probability accuracy

	Generation Condition	Passive Condition
Delay	15.11 (12.47)	15.09 (13.8)
No Delay	12.62 (13.53)	11.49 (12.03)

Fig. 6
Mean difference scores for the crossed conditions, with standard deviations in parentheses

The means for the crossed conditions are presented in fig. 6. A factorial ANOVA showed a significant main effect for the delay manipulation, *F*(1, 1,656) = 24.42, p = .0001. The generation manipulation was not significant, *F*(1, 1,656) = .87, p = .3511, nor was the presentation x delay interaction, *F*(1, 1,656) = .81, p = .3673. Although no tests were run on the medians, the medians were somewhat higher for delay conditions (12 for delay/generation, 10 for delay/passive) than for non-delay (9 for no delay/generation, 10 for no delay/passive). This would appear to confirm the tentative findings from experiment 1: the test delay does decrease participants' accuracy in making probability judgment, although only by a small amount (about 3% in this case).

Given the literature on the generation effect, it was surprising that I did not find any significant impact on accuracy. I briefly considered whether there was an effect from generative congruence: that is, whether the marginals which were concerned in the generative statement were the same marginals concerned in the probability statement or not. A congruent statement would be 'There were more red marbles than blue marbles,' followed by a question about the probability of picking a blue marble. An incongruent statement would be ‘There were more red marbles than blue marbles,’ followed by a question about picking a red marble. Research on probability judgment shows that when participants are primed with inapplicable categories before making a probability judgment, they are less accurate than participants who were primed with applicable categories (Hanita, Gavanski and Fazio, 1997). This may be because participants are neglecting the categories which they were not interacting with in the priming phase. If something similar was happening in the generation effect in this study, participants would be performing worse in the incongruent generation condition and better in the congruent generation condition, masking the impact of generation on their accuracy.

	Generation Condition	Passive Condition
Congruent	14.29 (13.02)	14.37 (13.63)
Incongruent	14.9 (14.31)	14.12 (13.96)

Fig. 7
Mean difference scores for marginals in the generation and congruence conditions, with standard deviations in parentheses

A GLM procedure in SAS, however, revealed no significant main effect of generation, *F*(1, 648) = .19, p = .6643, or of congruence, *F*(1, 648) = .03, p = 868. There was also no significant interaction of congruence and the generation manipulation, *F*(1, 648) = .16, p = 6887. While it would still be useful to confirm these results in a balanced study design, the issue of congruent categorization would not appear to apply to this case. It seems more likely that there was some sort of ceiling effect occurring – since participants were already performing within 15% mean accuracy in all conditions when judging probabilities, there was not much room for them to improve.

Question type

Marginals	Conjunctions
10	10

Fig. 8.
Medians of the marginal and conjunction difference scores.

Difference scores were once again used to compare participant accuracy when judging marginal probability to their accuracy judging conjunction probability. The median difference scores for marginals and conjunctions, as seen in fig. 8, were exactly the same. The means were marginals (M = 13.62, SD = 13.35) and conjunctions (M = 13.54, SD = 12.77). There was no significant difference between these two means, *t*(439) = .09.

Conjunction/disjunction errors

Out of 756 trials that had a possibility of a conjunction/disjunction fallacy, 14 contained conjunction/disjunction errors, or an across-participants rate of about 2%. Out of the 14 errors recorded, 1 participant committed 3 errors, 2 participants committed 2 errors, 7 participants committed 1 error, and 11 participants committed no errors at all. 6 errors were committed in the generation condition, and 8 were committed in the passive condition – a relatively even split. 5 errors were committed in the delay condition, and 9 in the no delay condition. A matched-pairs randomization test on the number of errors participants committed in delay vs no-delay conditions showed the difference to be non-significant (P = .09).

V. Discussion

Taken together, the results are somewhat surprising. Participants had a very low error rate for the conjunction/disjunction fallacy (about 2% for both experiments), and on average they were within 15% of the objective probability in all conditions. Accuracy was negatively impacted by the two-minute delay in both experiments. But this difference was very small – a greater discrepancy score of about 3% - and there was no significant impact of delay on the number of conjunction errors. There was no significant effect of the presentation manipulation or of the generation effect. Finally, there was no significant difference between the accuracy scores for marginals or conjunctions in either of the two experiments.

This largely supports the null hypothesis for no difference between conditions, and contradicts my original hypotheses about the probability integration explanation for the conjunction/disjunction fallacy. Participants were just as accurate when estimating integrated marginal probabilities as when judging conjunction probabilities, and they were just as accurate in the high integration condition of a segmented presentation as in the less integrative condition of the whole presentation. Conjunction/disjunction errors were also unaffected by the presentation manipulation. While the delay manipulation did have some impact on accuracy, taken as a whole, the probability integration hypothesis cannot alone explain the conjunction/disjunction fallacy. Participants did not have any special difficulty dealing with probability integration in this task. It seems more likely that participants were experiencing forgetting in the delay condition which led them to be less accurate due to loss of exact information, rather than due to an enhanced integrative load.

While my own hypotheses were mostly disproved, I turn next to the other three explanations for the conjunction/disjunction fallacy. The numeracy explanation also fared poorly in these experiments. Participants were highly accurate in every condition, and had no apparent trouble converting the number of one marble or another into a probability estimate. They also made nearly no conjunction/disjunction errors. This was a difficult task, despite its comparative simplicity: participants had only six seconds to look at the marbles (pilot testing indicated that this was not enough time to count each type of marble), and then had to provide probability estimates for two dimensions that they were unable to predict. The fact that participants were able to perform the necessary mental calculations and produce accurate numerical probability estimates and avoid conjunction/disjunction errors would also seem to reject the numeracy hypothesis as a complete explanation for the conjunction/disjunction fallacy.

As stated above, these experiments did not contain a strict test of the semantic explanations for the conjunction/disjunction fallacy. But it would seem to be important that participants were not deterred by the use of the word ‘probability,’ rather than the frequency model that some researchers have suggested would reduce conjunction errors (Hertwig and Gigerenzer, 1999). A true test of the semantic explanations would have to compare accuracy and errors on this task between a probability model and a frequency model, but it is hard to imagine conjunction errors decreasing from the 2% rate found in both experiments.

The theory that fared the best in these experiments was the social explanation. In a task without any attached social heuristics or stereotypes, participants were highly accurate and did not commit any conjunction/disjunction fallacies. They were not affected by integration manipulations. This research, thus, can be seen as confirming Barch et al.’s (2012) findings – in a clearly set toy problem, participants are able to closely estimate marginal and conjunction probabilities. Apparently it is only in socially relevant problems that people begin to make conjunction/disjunction errors.

There are, of course, other possible explanations for why participants were so good at the marble judgment task. Participants were presented with all the information they needed in visible format before they were asked to judge probabilities. In previous studies that have tested non-socially relevant conjunction tasks, participants were given word problems (Gavanski and Roskos-Ewoldsen, 1991), or had to recall probabilities from memory (Costello, 2009a). It is possible that the low incidence of conjunction/disjunction errors on the marble task is part of a picture superiority effect – people tend to remember information if it is presented in visual format, rather than only in a word-based format (Nelson and Reed, 1976). The conditions in the marble task were certainly conducive to the picture superiority effect: the effect is strongest when participants are given a longer response time in which to recall information, and in this task participants were not given a response deadline but were allowed to take as long as they liked. It is possible that a socially-relevant task that used a visual format like the marble task would also demonstrate high accuracy and low conjunction errors among participants.

VI. Future directions

More research needs to be done before any final conclusions can be reached about the conjunction/disjunction fallacy. A good extension of the marble task would be to create a visual socially-relevant task and see if there is a decrease in accuracy and an increase in errors relative to the marble task. Barch and colleagues are currently designing and conducting experiments in this vein (Barch et al., 2012). These include experiments about whether different types of flesh-colored marbles might activate social heuristics and cause an increase in conjunction errors, and experiments using visuals of human faces to activate the socially-relevant categories and conjunctions of race and occupation or race and perceptions of criminal status. If it is found that participants commit conjunction or disjunction errors in tasks that are both visual and socially-relevant, this would support the social explanations of the conjunction/disjunction fallacy. If, however, participants continue to commit relatively few errors, it would be more likely that the visual task is somehow cognitively easier or simpler than the word problems used by most conjunction researchers. It would also be possible to reverse this line of research and see how participants perform when they are given the toy marble task in a word problem format compared to a visual format. If participants make conjunction errors in the word-problem format, this would count against the social explanations of the conjunction/disjunction fallacy. If, however, participants remain relatively accurate in the word-problem condition, this would suggest that the visual task is not part of the explanation for the conjunction/disjunction fallacy.

Another important question raised in this paper is why I did not find a generation effect in experiment 2. In the results section above, I argue that this is probably because of a ceiling effect, since participants were already highly accurate when making probability judgments in the marble task. If this is the case, it would be interesting to apply the generation manipulation to cases which tend to have low accuracy and high incidence of conjunction errors, such as the traditional Linda task. For instance, if participants are given a socially-relevant visual task, as described above, and are then asked to generate information about what they just saw, they may be more accurate than participants in a passive viewing condition.

Finally, applying the delay manipulation to other conjunction/disjunction tasks could also reveal more about the processes underlying the conjunction/disjunction fallacy. In experiment 1 and experiment 2, I found a significant but small negative effect of delay on participant accuracy. However, as I also showed, performance on average in the marble task was highly accurate. In a socially-relevant task where accuracy is generally lower, a delay might have more of an impact on participant performance. If participants are already having trouble keeping socially-relevant categories straight, a delay could exacerbate those problems by reducing exact, verbatim knowledge and encouraging participants to rely more on their ingrained social heuristics and stereotypes in order to fill the gap left by forgetting. If delay has more of an impact in socially-relevant visual tasks than in the marble task, this could support the social explanations for the conjunction/disjunction fallacy.

VII. Conclusions

The experiments in this paper showed that in a visual, non-socially-relevant task, participants were highly accurate in judging probabilities and made very few conjunction errors. Participants were also equally accurate when judging conjunction and marginal probabilities. This finding would seem to confirm the social explanations for the conjunction/disjunction fallacy: in a task that did not activate any social heuristics, participants did not make any errors. This also discounts the numeracy and probability integration explanations for the conjunction/disjunction fallacy: participants were highly accurate about marble probabilities, and did not perform worse in high integration conditions, except for in the delay condition. Overall, while more research remains to be done, these results are strongly in favor of Tversky and Kahneman's (1983) original theory. The conjunction/disjunction fallacy would appear to be a function of social heuristics and stereotypes, and not a function of inherent difficulties or errors made when dealing with pure probability.

Works Cited

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology, 65*(6), 1119-1131. doi:10.1037/0022-3514.65.6.1119

Barch, D.H., Schultz, J., Chechile, R.A., & Sommers, S.R. (2012). Perceptual priming reduces conjunction errors: Probabilistic reasoning with social and non-social stimuli. Paper presented at the 83rd Annual Meeting of the Eastern Psychological Association, Pittsburgh, PA.

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language, 19*(2), 199-210. doi:10.1111/j.1468-0017.2004.00254.x

Chechile, R. A., & Soraci, S. A. (1999). Evidence for a multiple-process account of the generation effect. *Memory, 7*(4), 483-508. doi:10.1080/741944921

Costello, F. J. (2009). Fallacies in probability judgments for conjunctions and disjunctions of everyday events. *Journal of Behavioral Decision Making, 22*(3), 235-251. doi:10.1002/bdm.623

Costello, F. J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making, 22*(3), 213-234. doi:10.1002/bdm.618

Crisp, A. K., & Feeney, A. (2009). Causal conjunction fallacies: The roles of causal strength and mental resources. *The Quarterly Journal of Experimental Psychology, 62*(12), 2320-2337. doi:10.1080/17470210902783638

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning, 14*(2), 182-199. doi:10.1080/13546780701643406

Defeyter, M. A., Russo, R., & McPartlin, P. L. (2009). The picture superiority effect in recognition memory: A developmental study using the response signal procedure. *Cognitive Development, 24*(3), 265-273. doi:10.1016/j.cogdev.2009.05.002

Dougherty, M. R., & Sprenger, A. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. *Journal of Experimental Psychology: General, 135*(2), 262-281. doi:10.1037/0096-3445.135.2.262

Doyle, J. K. (1997). Judging cumulative risk. *Journal of Applied Social Psychology, 27*(6), 500-524. doi:10.1111/j.1559-1816.1997.tb00644.x

Fantino, E., & Stolarz-Fantino, S. (2007). Enhancing sensitivity to base-rates: Natural frequencies are not enough. *Behavioral and Brain Sciences, 30*(3), 262-263.

Fisk, J. E., & Pidgeon, N. (1997). The conjunction fallacy: The case for the existence of competing heuristic strategies. *British Journal of Psychology, 88*(1), 1-27. doi:10.1111/j.2044-8295.1997.tb02617.x

Fisk, J. E., & Pidgeon, N. (1998). Conditional probabilities, potential surprise, and the conjunction fallacy. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 51A*(3), 655-681. doi:10.1080/027249898391576

Gardiner, J. M. (1988). Generation and priming effects in word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 495-501. doi:10.1037/0278-7393.14.3.495

Gardiner, J. M., & Rowley, J. M. (1984). A generation effect with numbers rather than words. *Memory & Cognition, 12*(5), 443-445.

Gavanski, I., & Roskos-Ewoldsen, D. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology, 61*(2), 181-194. doi:10.1037/0022-3514.61.2.181

Hanita, M., Gavanski, I., & Fazio, R. H. (1997). Influencing probability judgments by manipulating the accessibility of sample spaces. *Personality and Social Psychology Bulletin, 23*(8), 801-813. doi:10.1177/0146167297238002

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12*(4), 275-305. doi:10.1002/(SICI)1099-0771(199912)12:4<275::AID-BDM323>3.0.CO;2-M

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 180-190. doi:10.1037/0278-7393.14.1.180

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454. doi:10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251. doi:10.1037/h0034747

Kinjo, H., & Snodgrass, J. G. (2000). Does the generation effect occur for pictures? *The American Journal of Psychology, 113*(1), 95-121. doi:10.2307/1423462

Kutzner, F., Freytag, P., Vogel, T., & Fiedler, K. (2008). Base-rate neglect as a function of base rates in probabilistic contingency learning. *Journal of the Experimental Analysis of Behavior, 90*(1), 23-32. doi:10.1901/jeab.2008.90-23

Lechuga, J., & Wiebe, J. S. (2011). Culture and probability judgment accuracy: The influence of holistic reasoning. *Journal of Cross-Cultural Psychology, 42*(6), 1054-1065. doi:10.1177/0022022111407914

McCloy, R., Byrne, R. M. J., & Johnson-Laird, P. (2010). Understanding cumulative risk. *The Quarterly Journal of Experimental Psychology, 63*(3), 499-515. doi:10.1080/17470210903024784

Miller, P. H., & Bjorklund, D. F. (1998). Contemplating fuzzy-trace theory: The gist of it. *Journal of Experimental Child Psychology, 71*(2), 184-193. doi:10.1006/jecp.1998.2471

Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin, 10*(2), 243-252. doi:10.1177/0146167284102010

Nelson, D. L., Reed, V. S., & Walling, J. R. (1976). Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 523-528. doi:10.1037/0278-7393.2.5.523

Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. *Journal of Behavioral Decision Making, 21*(4), 471-490. doi:10.1002/bdm.615

O'Leary, Daniel E. (2011). The emergence of individual knowledge in a group setting: Mitigating cognitive fallacies. *Group Decision and Negotiation, 20*(1), 3-18. doi:10.1007/s10726-010-9201-y

Peynırcıoğlu, Z. F. (1989). The generation effect with pictures and nonsense figures. *Acta Psychologica, 70*(2), 153-160. doi:10.1016/0001-6918(89)90018-8

Reyna, V. F., Nelson, W. L., Han, P. K. & Dieckmann, N. F (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943-973. doi: 10.1037/a0017327

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

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592-604. doi:10.1037/0278-7393.4.6.592

Soraci, S. A., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 67-78. doi:10.1037/0278-7393.20.1.67

Stolarz-Fantino, S., Fantino, E., & Van Borst, N. (2006). Use of base rates and case cue information in making likelihood estimates. *Memory & Cognition, 34*(3), 603-618.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232. doi:10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315. doi:10.1037/0033-295X.90.4.293

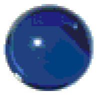
Wolfe, C. R., & Reyna, V. F. (2010). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making, 23*(2), 203-223. doi:10.1002/bdm.650

Wolford, G., Taylor, H. A., & Beck, J. R. (1990). The conjunction fallacy? *Memory & Cognition, 18*(1), 47-53.

Appendix I: materials from experiment 1

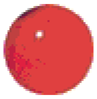
Marble types

Blue Striped



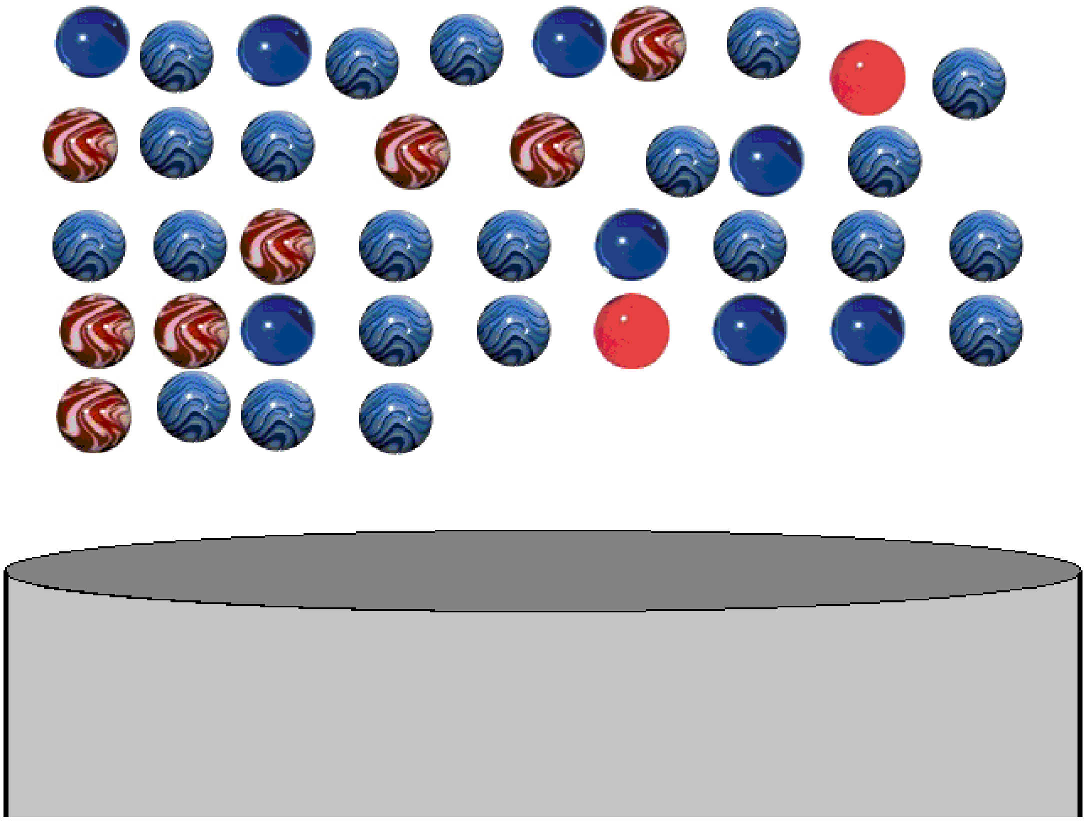
Blue Solid

Red Striped

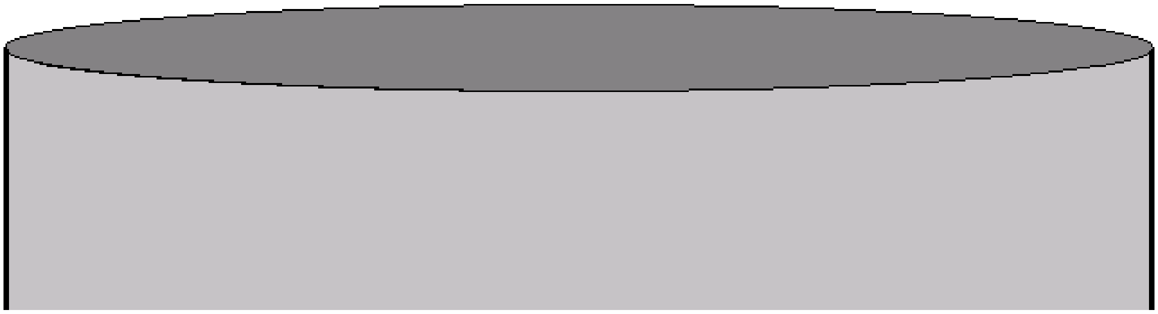


Red Solid

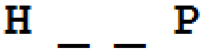
Whole marble presentation:



Screen after marbles fall into bucket:

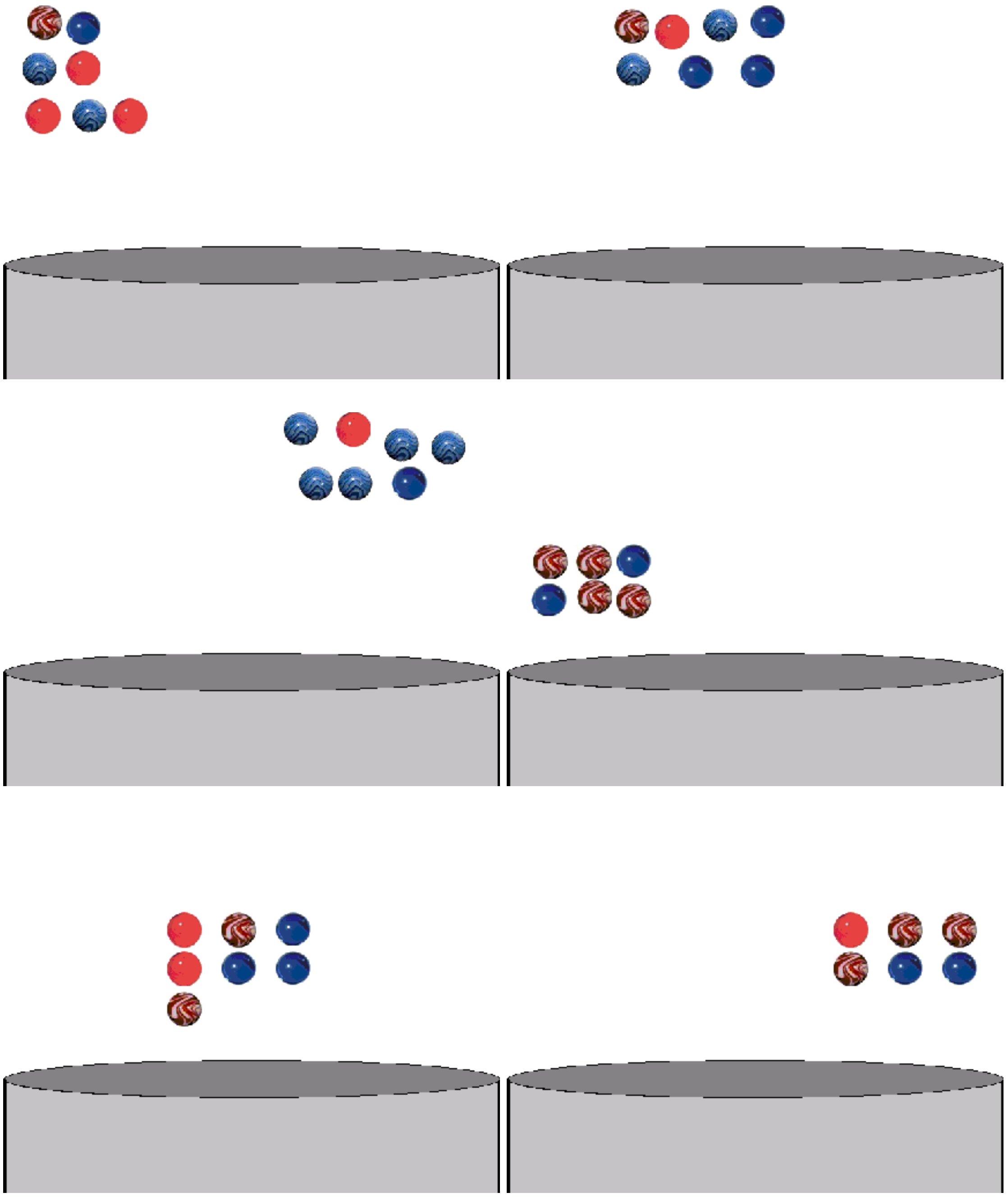


Word-fill example:



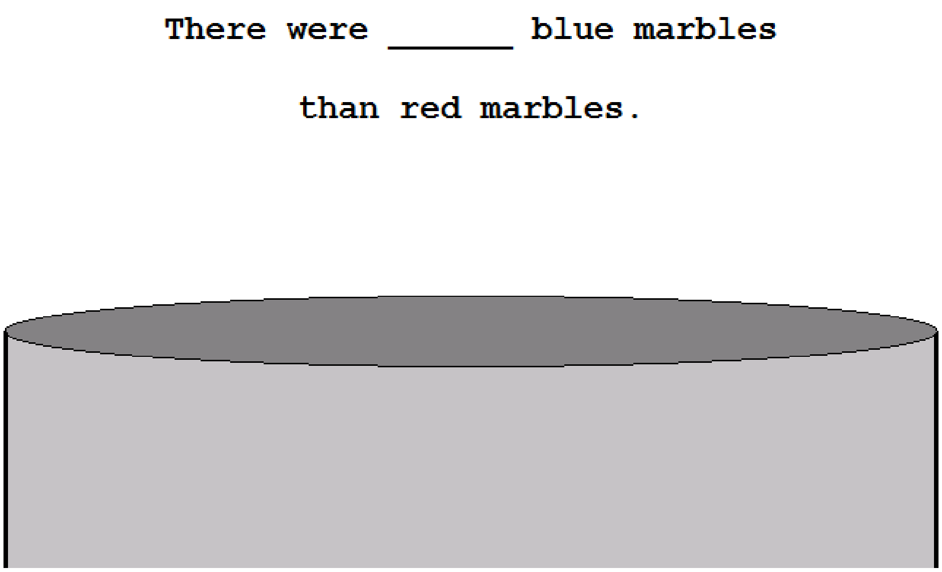
H _ _ P

Segmented presentation:



Appendix II: materials from experiment 2.

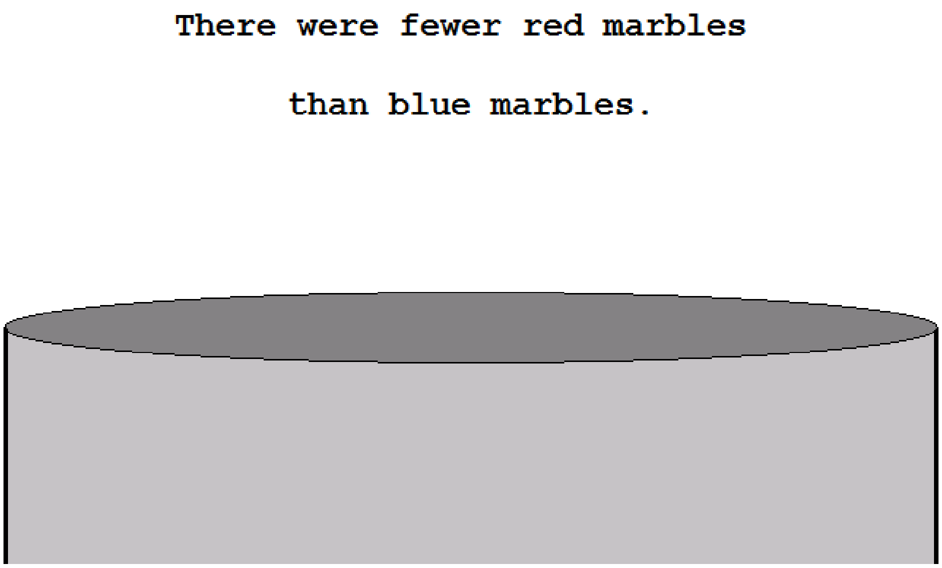
Generation sentence:



There were ______ blue marbles

than red marbles.

Passive statement:



There were fewer red marbles

than blue marbles.