Judgment and Decision Making, Vol. 17, No. 4, July 2022, pp. 797–815

The effect of a reference point in task difficulty: How does a task that becomes irrelevant affect effort, feelings and perceptions

Alisa Voslinsky* Ofer H. Azar†

Abstract

We examine the effect of an irrelevant task that may become a reference point on subjects' effort, feelings and perceptions. All subjects complete up to 25 tasks and are paid $0.10 per task solved correctly. However, some subjects have an easy task of finding one letter and others have a hard task of finding two letters. In the irrelevant-task treatment conditions subjects are told about the two types of tasks and are then assigned randomly to one. In addition, there are two control conditions, and in each control condition subjects are assigned to a specific task without the other task being possible or mentioned. Subjects in the irrelevant-task treatments express more positive (negative) feelings when assigned to the easy (hard) task. The control conditions that have no reference point of another task are in between the two irrelevant-task treatments in the feeling ratings. We hypothesized that for a given task, the subjects in the experimental conditions that have more positive feelings will also solve more tasks, but this hypothesis was not supported by the data. Finally, subjects who receive the easy task complete more tasks than the ones with the hard task.

Keywords: irrelevant task; real-effort task; effort; reference point; incentives; feelings and perceptions

*Corresponding author. Department of Industrial Engineering and Management, Sami Shamoon Academic College of Engineering, Ashdod, Israel. E-mail: alisavo@ac.sce.ac.il. https://orcid.org/0000-0002-5545-4160.

†Department of Business Administration, Ben-Gurion University of the Negev, Beer Sheva, Israel. Email: azar@bgu.ac.il. https://orcid.org/0000-0003-0154-327X.

This study was supported by a grant from Ben-Gurion University of the Negev. We thank three anonymous referees for helpful comments.

All data are available at https://drive.google.com/file/d/1M8mC2B6TbJRfTH5qlqULBVxkTj0lxh3m/view?usp=sharing.

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

1 Introduction

One of the central issues in motivating human behavior is incentives. People respond to the incentives they face. In particular, if one can exert effort and get rewarded for it, his decision about the level of effort should depend on how effort translates into rewards, which also means the relationship between cost and benefit. We study here whether effort is also affected by psychological manipulations that do not affect the economic incentives. In particular, we explore a psychological manipulation of creating a reference point for the difficulty level of the task.

We provide the subjects with two alternative tasks that could be chosen for them, and then one of the two becomes the actual task they can perform in order to earn money, and the other becomes irrelevant from an economic perspective – but possibly is relevant from a psychological perspective as it may create a reference point for the task difficulty. Subjects are told that the task can be to find one letter or to find two letters (in given positions on a page full of letters). The task is repeated 25 times, where the subject can choose how many of these to attempt and is paid according to the number of tasks solved correctly. The payment scheme is mentioned in the initial stage. Some of the subjects are given the one-letter task whereas other subjects are given the two-letter task, but in both cases the payment per correct task is the same; this means that in the two-letter task, one has to make about twice the effort to earn the same compensation. We also have two control groups, where subjects have to find either one or two letters in each task, but without the preliminary stage telling them that these two options exist and they will be allocated randomly to one of them. At the end, subjects are asked to complete a few short questions about their feelings and perceptions, and a few general demographics questions. The four feelings questions are measured on the scales of disappointed-satisfied, upset-pleased, sad-happy, and angry-calm. The two elicited perceptions about the task are on the scales of boring-interesting and difficult-easy.

We can measure effort by the number of tasks attempted or by the score (i.e., the number of tasks solved correctly). Our design allows us to compare how the difficulty of the task (one or two letters to find) affects effort, feelings and perceptions; and it also allows us to examine the effect of the reference point. The two control groups do not have a reference point, because they receive only their own version without being told about other conditions with different tasks. In the irrelevant-task treatment, however, some subjects get the one-letter task but have a reference point of a two-letter task (which they know they had an equal chance to be assigned to), whereas the other subjects get the two-letter task but have a reference point of the one-letter task. Consequently, in the irrelevant-task treatment, some have a reference point of a harder task whereas others have a reference point of an easier task.

Our study contributes to the literature about psychological manipulations that could affect feelings and effort levels, and have potential implications for several real-life areas. For instance, understanding how the existence of an alternative task may create a reference

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

point and may affect feelings and effort could be relevant to managers in the context of designing incentives in order to improve workers' motivation and performance. Such understanding can also be useful to teachers who want to motivate their students to improve their school performance, and to parents who want to motivate their children in certain areas.

The situation of tasks that pay the same but have a different difficulty level, where the worker does not know early on which task he will receive, happens sometimes in the real world. Some large employers such as governmental agencies or big firms recruit new workers for several positions without specifying from the very beginning which position each worker will get. Often the new workers have joint training at least in the initial stage, and only at an advanced stage workers are divided to positions that are not homogenous in their difficulty level, but are paid the same. For example, in a bank some workers may be assigned to a difficult position where there is a constant queue of customers whereas others are assigned to a position with lighter workload. Alternatively, some workers may be assigned to branches that are convenient to reach and others to more remote branches, or ones without dedicated employees' parking, etc. A similar situation can arise when new workers in the tax authority are assigned to different positions that pay the same but have different tasks with different difficulty levels. In the army it is also common that new soldiers do not know in advance what exactly they will do, and later are assigned to positions that differ in difficulty, but are being paid the same. In a restaurant some waiters get tables near the kitchen so their trips from the kitchen to the tables are short, whereas others get tables that are more remote (maybe even on a second floor) and require much more effort. Bus drivers can get convenient routes in empty and scenic roads outside the city, or get congested and tiring routes in the city center. Salespeople may get the same salary and same commission per sale, but some deal with customers who are harder to sell to, or are assigned to areas that are more remote and require more driving, etc. All these situations correspond to our study: a worker knows that tasks with several difficulty levels are possible and all pay the same, he later finds out to which exact task he is assigned, and then he chooses his effort level. The question is whether the effort level chosen by the worker is affected by the task that became irrelevant (i.e., the task that was possible but was not chosen for this worker eventually), because it may create a reference point and affect the worker's feelings, which in turn affect his behavior.

1.1 Incentives and effort

The effect of monetary incentives on effort was widely studied, and economists suggest that a higher payment provided for a task increases performance. Gneezy and Rustichini (2000) perform experiments in which they pay subjects for a set of 50 questions taken from an IQ test. The first group is paid a fixed amount, and three other groups are paid per each question that they answer correctly, with the payment per correct answer varying between the groups. Gneezy and Rustichini find that compensation is positively correlated with

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

performance. Other studies find a similar tendency (e.g., Rivas & Sutter, 2008; Clark et al., 2010; and Gächter & Thöni, 2010). Following these findings, we design an experiment where subjects are paid per correct answer, in addition to a fixed amount. This payment scheme allows us to measure the effort that subjects are willing to exert, by counting the correct answers and the attempted tasks for each subject.

1.2 Comparison income and effort

Some studies consider income comparisons and labor market outcomes, attempting to understand the effect of the income comparisons on performance. Clark et al. (2010) analyze the effect of income comparisons on effort using both experimental and survey data. They use the standard gift-exchange game (Fehr et al., 1993), where subjects that play as employers choose incomes, and the effect of income comparisons is derived from individuals' observed effort decisions. Clark et al. show that individual effort depends on both one's own income and the individual's position in the relevant income distribution. That is, subjects compare their own payment with others, and those who are paid up relatively well make more effort. Gächter and Thöni (2010) conduct three-person gift-exchange experiments using the strategy method. In their experiment, an employer is matched with two workers, which allows testing workers' reactions to wage inequality. Subjects are asked to choose their effort in reaction to various hypothetical income distributions. Gächter and Thöni find that employees' effort is positively correlated to their own wages, and the majority of individuals reduce their effort when faced with disadvantageous wage discrimination. However, advantageous wage discrimination does not increase the effort on average. Further, Rivas and Sutter (2008) conduct an experiment where one principal is matched with four workers, so the principal can choose unequal wages for the workers. The data show that higher workers' wage increases their efforts, but their efforts are negatively affected by high levels of wage inequality, i.e., when other workers' wage is higher than theirs. In addition, Bracha et al. (2015) experimentally test whether relative pay affects labor supply. They define relative pay as the worker's current pay relative to his past wages, or the worker's wage relative to other workers' wages. Bracha et al. find that when subjects' wage is lower than others' wage, they supply significantly less work time relative to subjects whose wage is relatively high. Moreover, the relatively low-paid subjects supply significantly more work time when they are unaware of the higher pay rates. Further, the data show that when offering individuals a higher wage than they had previously received, they work more than individuals who were offered less pay than they had previously received.

Additional studies test the effect of different wages on effort. For example, Burchett and Willoughby (2004) perform an experiment where they offer different monetary compensation for identical tasks and provide different information about alternative compensation systems. Their data show that work productivity varies significantly only when workers are aware of the alternative compensation systems. Workers who are paid the flat rate significantly decrease their work productivity, whereas those who are paid per produced

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

unit significantly increase their productivity. Greiner et al. (2011) conduct an experiment using a real-effort task. They pay equal wages for the first part of the experiment, but different wages in the second part, which is known to workers in one version and unknown in another. Greiner et al. conclude that when subjects are not aware of their peers' wages then their own wage change does not affect performance. However, when peers' wages are public, then higher-wage workers increase their effort, and the lower-wage workers decrease their effort.

As these studies show, people tend to compare their income to others' income, which in turn affects their effort in a specific task, versus the case when others' wages are unknown. We ask whether this effect holds when paying equal wages for different tasks. Particularly, whether comparing their own task to the alternative task, more or less difficult but with equal wage, affects subjects' effort versus the case of being unaware of the alternative task.

1.3 Reference dependence and effort

When outcomes are below the reference point people perceive them as a loss but perceive them as gains when the outcome is above the reference point (Kahneman & Tversky, 1979). Tversky and Kahneman (1991; 1992) present in their studies a formal analysis of the reference-dependence effect on effort and consider the change in preferences based on realized gains or losses. Following the above, a question arises when discussing the theory of reference-dependent preferences (Farber, 2008): what actually defines the reference point?

Evidence from previous studies indicates that expectations could serve as a reference point (e.g., Shalev, 2000; Kőszegi & Rabin, 2006, 2007, 2009). The expectations of people affect their feelings in real situations, which in turn affect their reactions. Abeler et al. (2011) use a real-effort experiment in which they manipulate the rational expectations of subjects and test subjects' effort in a tedious task. They find that subjects expecting high compensation work longer and earn more money than in the opposite case. Moreover, an unexpected assignment to different tasks could lead to a different effort exertion. Bushong and Gagnon-Bartsch (2020) examine experimentally the effect of positive and negative surprise on willingness to work. In the first part of the experiment, subjects practice one task out of two, one more onerous than the other. One group is assigned their task by chance just before working, while another group knew in advance which task they would face. Hours later, in the second part, subjects work on that task. Participants assigned to the less-onerous task by chance are more willing to work, while participants assigned to the more-onerous task by chance are less willing to work, compared to subjects who knew for certain which task they would face. These results show that positive and negative surprise affects subjects' willingness to work.

Furthermore, past self-achievements can be used by people as a reference point. Burdina and Hiller (2018) use empirical marathon data of past running history for runners and

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

investigate how distance from reference points affects future performance. They find that personal bests are served as reference points and affect the future performance of runners.

Another candidate to serve as a reference point is peer performance. Several field studies report generally positive performance effects when providing subjects relative performance information (Azmat & Iriberri, 2010; Chen et al., 2010; Allcott, 2011; and Allcott & Rogers, 2014). Additionally, Eyring and Narayanan (2018) conduct a field experiment where they provide to online students two reference points for peer-performance comparison. They find that when providing to students the top quartile reference point to compare themselves to, those who perform below the median decrease their performance, but students who perform above the median increase their performance, relative to students provided the median as a reference point.

We also study the possible effect of a reference point, but address a potential reference point that was not explored before. In particular, we study whether an alternative task that was presented as a possible task but then was not chosen and should therefore be irrelevant from an economic perspective, could serve as a reference point and affect subjects' effort. We hypothesize that knowing that another task that is either more difficult or easier could be chosen, may affect the subjects' feelings, and consequently also their choice of effort. To study this, we design an experiment that attempts to test whether subjects change their effort because they compare their own task, which is randomly chosen for them, to another task they could have. In particular, we ask whether receiving an easier (harder) task, which means that the alternative task that could be chosen is more (less) difficult, could be perceived as a gain (loss) and result in more (less) positive feelings and more (less) effort.

2 Experimental design and research hypotheses

2.1 Participants

We conducted an online experiment with 330 participants, using the Qualtrics software. Subjects were recruited using the CloudResearch platform (Litman et al., 2017). CloudResearch, formerly TurkPrime, is a participant-sourcing platform for online research and surveys, which provides access to diverse, high-quality respondents around the world. The subjects are registered on the platform of Amazon Mechanical Turk (known also as MTurk). We contacted English-speaking respondents living in the United States only. To obtain high quality of the subject pool we only allowed registration to our study by subjects who completed at least 100 previous tasks on MTurk. We also required that they had at least 99% approval rate in previous tasks, i.e., that in at least 99% of their previous tasks, the person who offered the task found their performance satisfactory and approved them for payment. No individual participated in more than one version of the experiment. 46% of the participants were female, and the age ranged between 19 and 84 (M = 40.45, SD = 11.64).

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

2.2 The Experiment

The experiment contained three sections (the experimental materials are included in the Appendix). The first section consisted of the real-effort task of finding letters on pages (Azar, 2019). This task involves several pages filled with letters, where the subject is asked to find which letter appears in a specific page number, line number and position number in the line. This fits the purpose of a simple and quick task that does not require complex instructions or practice and can be done online with many repetitions. The subjects were only allowed to participate using a computer (i.e., not a smartphone or a tablet). They accessed a file with pages that include letters that become smaller from page to page, increasing the difficulty of the task.

The subjects could attempt any number of tasks that they wanted, up to the maximum of 25 tasks. Because the difficulty level of the questions increases as one proceeds (the tasks proceed from the first pages with larger letters to the next pages with smaller and smaller letters), it was expected that subjects would start solving a few questions, but different subjects would finish solving the questions at different points, thus creating heterogeneity in how many questions were attempted and how many were solved correctly. Subjects did not get any feedback about the correctness of their answers, except after the experiment when they were paid based on the number of tasks they completed correctly.

In the treatment group, all subjects were told about the two possible tasks, from which one was going to be randomly selected for them. They were presented with the task of finding one letter (an easier task) and the task of finding two letters (a harder task). After the tasks' description subjects had to answer two example questions, one for every task type. They needed to solve the example questions correctly in order to proceed (they could try again if they did not succeed initially). Then subjects were randomly assigned to one of the two tasks. Consequently, some subjects faced 25 tasks of finding one letter, knowing that they could instead face 25 tasks of finding two letters.[1] For them, the task that was a possibility that became irrelevant and may become a reference point (finding two letters) was harder than their actual task, so we can refer to this treatment as 1LHIT (1L for finding one letter, and HIT for having a reference point of a Harder Irrelevant Task). The other subjects faced tasks of finding two letters knowing that the other possibility that became irrelevant was tasks of finding one letter, so they are denoted the 2LEIT (2 letters, Easier Irrelevant Task) treatment.

In the control group, all subjects were given one task only and they were not told about other conditions with other tasks. Half of the subjects received the one-letter task and the other half received the two-letter task. After the task's description subjects had to answer an example question, according to the task type they were assigned to. They needed to solve the example question correctly in order to proceed. After that, subjects proceeded to perform the task of finding one letter or two letters, according to their version. Therefore, all

[1]Each of the 25 tasks had five different versions (of similar difficulty) that were given randomly to subjects, to avoid a situation in which one subject could inform others online about the correct answers.

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

subjects performed only one task type, and they could complete up to 25 tasks of this type, without having a reference point of another task. We therefore refer to these conditions as 1LCtrl (1 letter, control group) for those whose task was to find one letter and 2LCtrl for the group with a two-letter task. The experiment was done between-subjects, i.e., every subject participated in only one of the four conditions.

For every task solved correctly (finding one or two letters correctly depending on the condition) subjects earned \$0.10. This amount was chosen based on the time it takes to solve questions and the common payments on similar online platforms. The results, with heterogeneity in the number of tasks attempted and with the average number being somewhere in the middle of the possible range of 0-25, show that the payment per task was chosen well.

In the second section, participants were asked to answer four questions on a semantic differential scale with nine options, concerning their feelings based on the task and condition they received. They were asked "Each rating scale consists of two opposite words with nine check boxes in between. Check a box between each pair of words that best describes how you feel about XXX", where XXX is different for each condition. For example, in the 1LCtrl condition it was "the task that requires to find one letter to get rewarded with a correct answer" and in the 1LHIT condition it was "being assigned to the task that requires to find one letter to get rewarded with a correct answer, rather than to the task that requires to find two letters." The four scales that followed captured feelings and were disappointed-satisfied, upset-pleased, sad-happy, and angry-calm. These scales were chosen to capture the most relevant emotions and feelings in the experimental situation of knowing that two treatments were possible, one harder than the other, and the subject was assigned randomly to one of them. Then subjects in all conditions were asked "The task of finding the letters was:" followed by scales with nine points for two perceptions of the task, boring-interesting and difficult-easy. Notice that in all six questions, 1 is the extreme bad (e.g., disappointed or boring) and 9 is the extreme good.

Finally, in the third section, some basic demographic data (e.g., gender, age, years of schooling and number of university-level courses in economics taken) were collected. The experiment lasted 18.5 minutes on average (SD = 13.8). Every subject who answered all the questions in the second and third sections received a basic compensation of \$0.5 in addition to her earnings for the tasks solved in the first section. All participants knew the size of the basic compensation and the bonus for every correct answer in advance. In this way, the subjects have a real opportunity cost when staying and solving additional questions because they can leave the experiment earlier if they solve fewer questions.

2.3 Research hypotheses

When one knows that he could be assigned to either an easy task or a difficult task, with a clear preference for the easy task because it allows to make money more easily and the difficult task is not intellectually challenging or otherwise attractive, it is natural to expect

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

that subjects hope to be assigned to the easy task. Consequently, those who get the easy task should have more positive feelings about their task assignment than those who get the hard task. Therefore, our first hypothesis is as follows:

H1*: In the irrelevant-task treatment conditions, scores given to the four feelings questions will be higher (more positive) in the 1LHIT condition than in the 2LEIT condition.*

We also expect the perception of the task to be more negative (more boring and more difficult) for those who were assigned the hard task:

H2*: In the irrelevant-task treatment conditions, scores given to the two perception questions will be higher (more positive) in the 1LHIT condition than in the 2LEIT condition.*

In the two control conditions, there is no reference point of another task difficulty that could be chosen for the subject. Therefore, we do not expect the subjects to differ in their feelings based on their assignment to the task in the control conditions and we hypothesize as follows:

H3*: In the two control conditions (1LCtrl and 2LCtrl), scores given to the four feelings questions will be similar on average.*

Because we expect those in the 1LHIT condition to develop positive feelings and those in the 2LEIT condition to develop negative feelings, whereas the control conditions subjects are not expected to develop strong feelings to either side, we expect the feelings in the control conditions to be rated between the ratings in 2LEIT and 1LHIT. For the sake of simplicity, we will compare each control condition to the irrelevant-task condition that has the same actual task (i.e., comparing 1LCtrl to 1LHIT and 2LCtrl to 2LEIT). That is, we hypothesize that:

H4*: In the control condition 1LCtrl, scores given to the four feelings questions will be lower than the scores given in the condition 1LHIT.*

H5*: In the control condition 2LCtrl, scores given to the four feelings questions will be higher than the scores given in the condition 2LEIT .*

Kahneman and Tversky (1979) show that a reference point affects people's decisions and they perceive losses or gains with regard to a reference point. Moreover, Tversky and Kahneman (1991; 1992) present the reference-dependence effect on effort. Further, various studies demonstrate that expectations as reference points affect reactions (Shalev, 2000; Kőszegi & Rabin, 2006, 2007, 2009), and factors that could serve as a reference point affect effort, such as expectations (Abeler et al., 2011), advance knowledge (Bushong & Gagnon-Bartsch, 2020), past self-achievements (Burdina & Hiller, 2018), and peer performance (Azmat & Iriberri, 2010; Chen et al., 2010; Allcott, 2011; and Allcott & Rogers, 2014). Following the above, we ask whether a task presented but not given to a subject could serve as a reference point that affects the effort he exerts. We hypothesize that having positive feelings after being assigned an easy task out of two possible tasks will result in more effort compared to solving the same task without knowing that a more difficult task was possible. This leads to the following hypothesis:

H6*: Effort in the 1LHIT condition will be higher than in the 1LCtrl condition.*

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

Similarly, we hypothesize that having negative feelings after being assigned a hard task out of two possible tasks will result in less effort compared to solving the same task without knowing that an easier task was possible:

***H7**: Effort in the 2LEIT condition will be lower than in the 2LCtrl condition.*

3 Results and discussion

We define two measures of effort. Answers is the number of tasks for which the subject provided answers (regardless of the correctness of the answers). Score is the number of tasks solved correctly. As explained earlier, a task can be finding one letter or two letters, depending on the relevant treatment. That is, Answers and Score refer to the number of tasks and not to the number of letters. Naturally, these two measures of effort are highly correlated (a correlation analysis gives $r = 0.98$), and we analyze both of them for the sake of examining the robustness of the results and not because they measure substantially different things. In Table 1 and Figures 1 and 2 we present the summary statistics of these variables and the six questions about feelings and perceptions, divided by the experimental condition.[2]

Table 1: Summary statistics.

Variable	Overall Mean	Overall St. Dev.	1LCtrl Mean	1LHIT Mean	2LCtrl Mean	2LEIT Mean
Score	9.95	7.95	13.71	11.19	8.16	7.00
Answers	10.97	8.67	14.95	12.46	8.90	7.84
Disappointed-satisfied	5.36	2.61	5.70	6.70	5.83	3.44
Upset-pleased	5.58	2.37	5.77	6.66	6.01	4.08
Sad-happy	5.55	2.28	5.66	6.56	5.93	4.21
Angry-calm	6.14	2.39	6.34	6.97	6.45	4.94
Boring-interesting	3.49	2.67	3.70	3.66	3.71	2.96
Difficult-easy	4.45	2.54	4.78	4.39	4.78	3.89
N	330		82	79	80	89

To test our hypotheses, we performed the t-tests for difference in means that compare between pairs of related conditions, for example between the two control conditions or between a treatment condition and a control condition with the same task. The results are summarized in Table 2.

[2]Balance tests of subject characteristics between treatments show no significant differences at the 5% level, except for number of economics courses taken, which is lower in the 2LEIT condition than in 2LCtrl condition (β = -.038, p-value = 0.033).

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

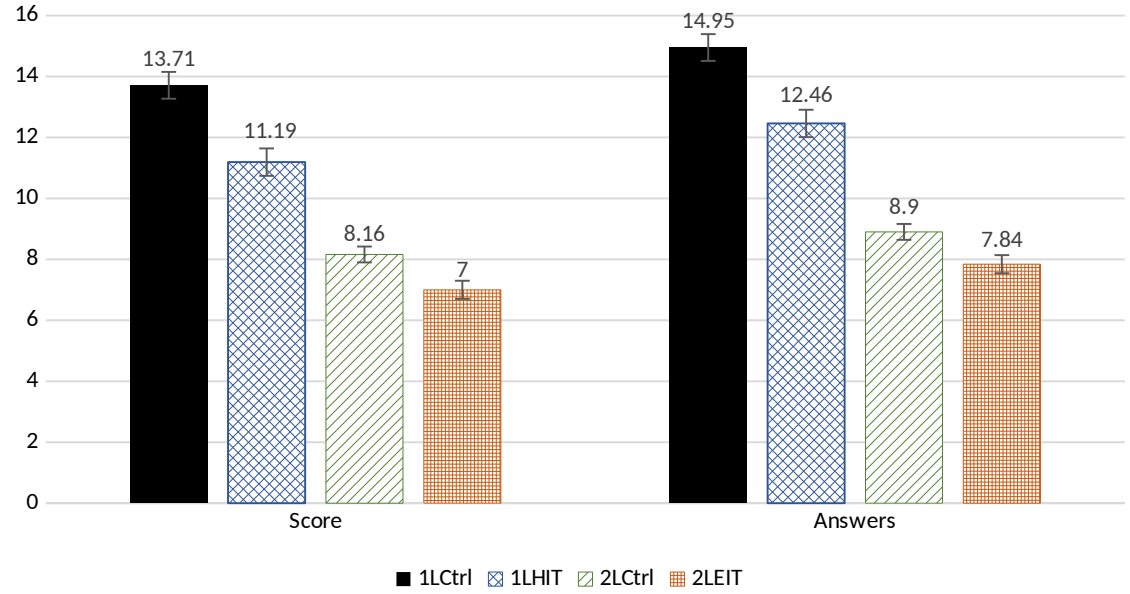

FIGURE 1: Means by experimental condition — Score and Answers. Error bars represent standard errors of the means. Answers is the number of tasks for which the subject provided answers (regardless of the correctness of the answers). Score is the number of tasks solved correctly. The conditions are as follows: 1LCtrl is one-letter task, 2LCtrl is two-letter task, 1LHIT is one-letter task with harder irrelevant task, and 2LEIT is two-letter task with easier irrelevant task.

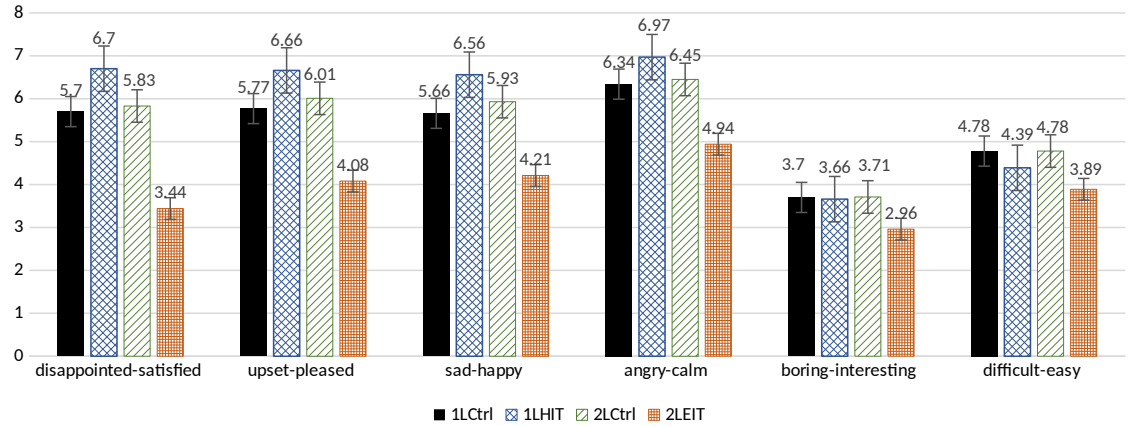

FIGURE 2: Means by experimental condition — feelings and perceptions. Error bars represent standard errors of the means. The conditions are as follows: 1LCtrl is one-letter task, 2LCtrl is two-letter task, 1LHIT is one-letter task with harder irrelevant task, and 2LEIT is two-letter task with easier irrelevant task.

First, it is worthwhile to examine as a validity test of our experimental design, whether the difficulty of the task has the expected effect. Because some subjects are required to find two letters to receive the $0.1 for a task whereas others are required to find only one letter for the same payment, based on economic logic we expect the number of tasks attempted

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

TABLE 2: *t*-tests for difference in means between the four conditions. P-values are 2-sided; 1-sided p-values would be these divided by 2.

		(1) 1LHIT–2LEIT	(2) 1LCtrl–2LCtrl	(3) 1LCtrl–1LHIT	(4) 2LCtrl–2LEIT
Score	Difference	4.1899	5.5448	2.5174	1.1625
	p-value	0.0002	0.0000	0.0568	0.2602
Answers	Difference	4.6130	6.0512	2.4955	1.0573
	p-value	0.0002	0.0000	0.0821	0.3522
Disappointed-satisfied	Difference	3.2580	–0.1299	–1.0011	2.3868
	p-value	0.0000	0.7177	0.0056	0.0000
Upset-pleased	Difference	2.5796	–0.2442	–0.8899	1.9339
	p-value	0.0000	0.4546	0.0118	0.0000
Sad-happy	Difference	2.3435	–0.2665	–0.8984	1.7115
	p-value	0.0000	0.4092	0.0070	0.0000
Angry-calm	Difference	2.0309	–0.1085	–0.6332	1.5062
	p-value	0.0000	0.7511	0.0717	0.0000
Boring-interesting	Difference	0.7032	–0.0174	0.0369	0.7574
	p-value	0.0844	0.9675	0.9315	0.0616
Difficult-easy	Difference	0.5048	0.0055	0.3881	0.8874
	p-value	0.2039	0.9888	0.3337	0.0233
N		168	162	161	169

and the number of tasks solved correctly to be higher when the task is easier. To test this, we examine the *t*-tests of Score and Answers in the second results column, which compares the two control conditions by considering the difference 1LCtrl–2LCtrl. We can see that indeed, the number of attempted tasks (Answers) and of correctly-solved tasks (Score) is substantially higher when the task requires to find one letter rather than two letters, with *p*-value of 0.0000. Subjects in the control condition with one-letter tasks attempted on average 6.05 more tasks (14.95 vs. 8.90, see Table 1) than in the control condition with two-letter tasks. In terms of tasks solved correctly the difference is again substantial (13.71 vs. 8.16). We can also consider the first column, which compares the two irrelevant-task treatments, and see once again that those whose task was to find one letter solved much more than those whose task was to solve two letters. In this comparison (the first column),

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

however, not only the task difficulty level plays a role but also the reference point (that does not exist in the control conditions) is different, and may affect behavior. The analysis above provides evidence that in this experiment the economic incentives function according to economic intuition: when the same payment is given for more time-consuming tasks, people choose to complete fewer of these tasks.

Next, let us consider the first hypothesis, about the feelings in the two irrelevant-task conditions. The *t*-tests that are relevant here are those reported in column (1) in Table 2, comparing between the 1LHIT and the 2LEIT conditions. The differences are based on the answers in the 1LHIT condition minus those in the 2LEIT condition, and higher values represent more positive values. Therefore, the positive values of these differences mean that among those who knew they may get either a hard or an easy task, the feelings of those who were assigned to the easy task are more positive; they are more satisfied, pleased, happy and calm, whereas those who were assigned the hard task are more disappointed, upset, sad and angry. The average difference between the two condition for these four feelings is between 2.03 (for angry-calm) to 3.26 (for disappointed-satisfied). All of the *t*-tests for difference in means related to the first hypothesis have a *p*-value of 0.0000.[3] Thus, our first hypothesis is strongly supported by the data.

Our second hypothesis is also about the comparison between the 1LHIT and the 2LEIT conditions (column (1) in Table 2), but here we consider the answers to the two questions that ask about the perceptions of the task. The results are in the predicted direction: those who received the one-letter task rate the task on average as more interesting and easier than those who got the two-letter task. However, in the difficult-easy scale the difference is not statistically significant, and in the boring-interesting scale it is statistically significant at the 5% level only in the 1-sided *t*-test.

Next, we hypothesized that in the two control conditions, because there is no reference point of another task difficulty that could be chosen for the subject, subjects will not differ in their feelings based on their task. Column (2), which compares between the two control conditions (1LCtrl and 2LCtrl), shows that as we hypothesized, the average scores given to the four feelings questions are similar in these two conditions. The difference in the average score for the four questions between the two control conditions is very small, between 0.11 and 0.27 and its 2-sided *p*-value is between 0.41–0.75.

It was less clear whether the ratings of the task perception would differ, because a task of finding two letters may be perceived as more difficult (or more boring) than a task of finding one letter, even without a reference point of the other task being possible. Therefore, we did not hypothesize about this comparison. However, the results allow us to look also at these two scores, and they show that these two ratings are essentially identical between the two control conditions. The average differences of these two ratings are less than 0.02 and the 2-sided *p*-values are 0.97–0.99.

[3] *p*-value = 0.0000 means that this is the computed value rounded to the fourth place after the decimal point, i.e., corresponding to stating that *p*-value < 0.00005.

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

We proceed to the fourth hypothesis and consider column (3), comparing 1LCtrl and 1LHIT. The values of the average differences of 1LCtrl-1LHIT for the four feelings are all negative (between –0.63 and –1), implying that among those whose task was to find one letter, the ones who knew they could be assigned to find two letters have more positive feelings (higher ratings) than those who had no such reference point of a harder task. This is consistent with our hypothesis. The 1-sided *p*-values are between 0.003 and 0.036. We did not make any hypotheses about the two perceptions regarding the task (the scales of Boring-interesting and Difficult-easy) because it is not clear that these should be affected by the irrelevant-task treatment, and indeed we can see that the difference in these ratings between the 1LCtrl and 1LHIT conditions is small and not statistically significant (*p*-values of 0.33–0.93).

The fifth hypothesis deals with the corresponding difference between the control and treatment conditions for those whose actual task was to find two letters. We can see in column (4) that the average value of 2LCtrl-2LEIT is positive for all four feelings (between 1.5 and 2.39). The difference between the two conditions is statistically significant with *p*-value = 0.0000 for all four feelings. This is in line with our hypothesis, that subjects who knew there are two possible tasks and were assigned to the difficult one, will experience more negative feelings than subjects who perform the same task without having a reference point of an easier task that could be chosen for them.

It is interesting to see that the absolute values of the differences here (1.5–2.39) are higher than those of 1LCtrl-1LHIT (0.63-1). We can consider the control condition as neutral (no reference point of another task existed), the 2LEIT condition as a loss (there is a reference point of an easier task and the subject "lost" and got the harder task), and the 1LHIT as a gain (there is a reference point of a harder task and the subject "gained" and got the easier task). Then the result that the absolute value of the difference of 2LCtrl–2LEIT is larger than that of 1LCtrl–1LHIT resembles the idea that losses loom larger than gains, captured in prospect theory (Kahneman & Tversky, 1979). Notice however that here we are considering ratings of feelings, following a task assignment to an easy or hard task, whereas the traditional findings about loss aversion refer to other contexts, such as gambles and loss or gain of money compared to a reference point, etc. In addition, if we consider the ratio in the absolute values of the differences (|2LCtrl–2LEIT|/|1LCtrl–1LHIT|), we have 2.39/1 = 2.39 (Disappointed-satisfied); 1.93/0.89 = 2.17 (Upset-pleased); 1.71/0.9 = 1.9 (Sad-happy); and 1.51/0.63 = 2.4 (Angry-calm). On average the ratio, which measures to what extent losses loom larger than gains in the feelings, is slightly above 2. This ratio of about 2 between the perception of losses versus gains also corresponds with a common finding in the literature about loss aversion related to money.

We turn now to the hypotheses that deal with effort. As discussed earlier we find evidence that the irrelevant-task treatments had the hypothesized effect on feelings. Will this effect on feelings translate also to behavior – choosing the effort level? Let us start with hypothesis 6, which considers the case of subjects who solved one-letter tasks. We

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

hypothesized that the positive feelings in the 1LHIT condition will translate into higher effort compared to the control condition. However, column (3) shows the opposite – the number of tasks attempted (Answers) and the number solved correctly (Score) are both higher in the 1LCtrl control condition than in the 1LHIT condition. The difference is not small – about 2.5 tasks in both variables – but its 2-sided p-value is not statistically significant at the 5% level.[4] Thus, hypothesis H6 is not supported by the data, which are in the opposite direction.

As for the seventh hypothesis, we consider column (4). Our hypothesis was that the negative feelings of those who were assigned the hard task knowing they could be assigned an easier task would lead them to exert less effort than in the control condition (subjects who solved two-letter tasks without a reference point of an easier task). The results are in the predicted direction, but are not statistically significant, with 1-sided p-values of 0.13-0.18.

To gain more insights about what affects effort in addition to the experimental condition, we ran several regressions. Some regressions analyze a pair of related experimental conditions, for example the two control conditions or the two treatment conditions (the conditions with an irrelevant task), or two conditions with the same task performed (e.g., the treatment condition and the control condition with the task of finding one letter). Some regressions analyze the entire data. To examine the robustness of the results we analyzed both measures of effort, Score and Answers, as possible dependent variables, although they are highly correlated. The results are summarized in Table 3 and Table 4 in an additional appendix at https://journal.sjdm.org/21/211201/robust.pdf.

4 Conclusion

This study contributes to the literature on psychological manipulations that may affect feelings and behavior, by examining subjects' feelings and effort exerted when they are paid equal wages for different tasks. In particular, we examine whether comparing their own task to an alternative irrelevant task, more or less difficult but with equal wage, affects subjects' feelings and effort exerted compared to the case of being unaware of the alternative task.

We use an online experiment to explore the effect of an irrelevant task that may become a reference point on subjects' effort, feelings and perceptions about the task. All subjects complete up to 25 tasks, as many as they wish, and are paid \$0.10 per correct answer. However, some subjects have an easy task of finding one letter and others have a hard task of finding two letters. The \$0.10 payment is for each correct task and not each correct letter, and therefore in the hard-task treatment subjects have to exert about twice the effort to earn the same reward compared to the easy-task treatment. In the irrelevant-task treatment conditions, subjects are told about the two types of tasks and are then assigned randomly to one of the tasks. In addition, there are two control conditions, where in each control

[4]Because the results are in the opposite direction to the hypothesis, we consider the 2-sided test despite having a hypothesis about a specific direction.

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

condition subjects are assigned to a specific task (finding one letter or two) without the other task being possible or mentioned. We hypothesized that those who were in the irrelevant-task treatments will express more positive feelings if they were assigned to the easy task and more negative feelings if they were assigned to the hard task. We also hypothesized that the control conditions who have no reference point of another task will be in between the two irrelevant-task treatments in the feelings, as they should not develop feelings from being assigned to a specific task when no other task was possible for them. Our hypotheses about the feelings were supported by the data.

We hypothesized that for a given task, the participants in the experimental conditions who have more positive feelings will also make more effort and solve more tasks, but this hypothesis was not supported by the data. In other words, although our experimental design (in the irrelevant-task conditions) of describing two tasks and assigning one of them achieved its expected impact on feelings, it did not produce a change in behavior in terms of choosing how many tasks to complete. In that decision, subjects followed their economic incentives without being influenced by their feelings.

The importance of the economic incentives is further observed when testing the effect of the task difficulty. We expected that the subjects who receive the easy task will complete more tasks than the ones with the hard task, given that the payment per task is the same. This follows from the economic logic that people should compare the cost and benefit of completing the tasks and therefore will choose to attempt more tasks when the cost is lower, given that the benefit is fixed. The data strongly supported this expectation.

The finding that the manipulation of introducing an irrelevant task had a strong effect on feelings and yet had no corresponding effect on behavior (effort level was not higher when feelings were more positive) is intriguing. A traditional economist may not find this so surprising, believing that indeed economic incentives are all that matter for the choice of effort and feelings are irrelevant. However, the literature in behavioral economics, economic psychology, judgment and decision making and other related fields shows that feelings do matter and that human decision makers often deviate from the assumptions of traditional economics about the fully-rational economic agent (sometimes referred to as homo economicus). We think that the explanation for the lack of effect of feelings on behavior in our study is the experience of the subjects. To obtain high quality of the subject pool we only allowed registration to our study by subjects who completed at least 100 previous tasks on MTurk (and had at least 99% approval rate in previous tasks). Some of these subjects may have completed far more than 100 previous tasks. This means that subjects were very experienced. As such, they learned to analyze effectively the connection between the time they need to spend on a task and the payment they receive, otherwise they could spend a long time on tasks that pay very little. As a result of this experience, they make informed decisions about how many tasks to complete, and their feelings do not affect their choice of effort. Recall that as our letters became smaller and smaller as the tasks progressed, the task became harder for the same payment, meaning that even if initially the

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

benefit of the payment for solving a task exceeded the time cost, this was likely to reverse at some point (at a harder task with smaller letters). Thus, we believe that our results show that people who are experienced in making a certain decision and received feedback about this decision many times in the past[5], learn to make informed decisions and ignore factors that are irrelevant from an economic perspective, such as their feelings. An interesting idea for future research is to conduct a similar experimental study but with subjects who are inexperienced and analyze to what extent the different subject pool will change the results. However, as many of the MTurk workers are highly experienced, this will probably require a different method of running the experiment in order to access inexperienced subjects.

Although the irrelevant task did not affect performance, we believe that our results suggest that thinking about how workers consider alternative tasks that they do not have to perform (the irrelevant task) is important for firms and other organizations. There are two reasons. First, because there is an effect of the irrelevant task on feelings, it could be that in other situations the feelings will also affect performance; as we mentioned above, our results may be related to the substantial experience that MTurk workers have in choosing which tasks to do and in evaluating costs versus benefits, but others may behave differently.

Second, both the firm and the workers can be better off if the firm creates better feelings for its workers by emphasizing and reminding the workers about a more difficult alternative task rather than an easier one. This is because the firm does not have to pay more, and still it gets workers who are happier, more satisfied and pleased, etc. Workers being happier may also have a positive externality on their co-workers because the entire work environment becomes more pleasant when people around are happy and pleased than when they are sad and upset. This could imply that also co-workers become more satisfied, and possibly it could also affect the performance of the co-workers. But even if the performance of the workers with the alternative task and of their co-workers is unchanged, the better feelings of the workers (and their co-workers) is beneficial for the firm, because being happier in the job results in lower chances that the workers will leave the firm, resulting in costs to the firm for recruiting and teaching new workers. Also, workers being unhappy in their jobs may require higher salaries to be retained than happy workers, again providing the firm an incentive to keep its workers happy.

Therefore, the study offers practical implications for firms and organizations, namely to do what is possible to make the workers think about the less pleasant jobs they could get and not the more pleasant ones. For example, suppose that in a fast-food restaurant the most attractive job is to be a seller who takes orders, the least attractive is to be a cleaner and in the middle is the job of preparing the food. Suppose that all workers receive the same salary. The restaurant has an incentive to remind the workers who prepare the food that they could have been the cleaners, which will make them happier than if they keep thinking they could have been the sellers. A similar point can be made in many other situations,

[5]For every task the subjects completed in MTurk in the past, they could see how much time they spent and how much they earned.

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

for example a building company where workers have tasks that differ in their difficulty and maybe also danger but are paid the same. In the governmental sector or not-for-profit organizations the same idea still applies. For example, there may be school inspectors for different neighborhoods, some of which are nicer to handle or maybe are closer to the office and therefore more convenient. The employer would want the workers to focus on the more difficult neighborhoods they could have received rather than on the more pleasant ones, so that they are happier about their job.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, *101*(2), 470–92.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9-10), 1082–1095.

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *The American Economic Review*, *104*(10), 3003-37.

Azar, O. H. (2019). Do fixed payments affect effort? Examining relative thinking in mixed compensation schemes. *Journal of Economic Psychology*, 70, 52–66.

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, *94*(7-8), 435–452.

Bracha, A., Gneezy, U., & Loewenstein, G. (2015). Relative pay and labor supply. *Journal of Labor Economics*, *33*(2), 297–315.

Burchett, R., & Willoughby, J. (2004). Work productivity when knowledge of different reward systems varies: Report from an economic experiment. *Journal of Economic Psychology*, *25*(5), 591-600.

Burdina, M. M., & Hiller, R. S. (2018). Performance as a Reference Point: Evidence from Marathon Data. *Scott, Performance as a Reference Point: Evidence from Marathon Data (December 12, 2018)*.

Bushong, B., & Gagnon-Bartsch, T. (2020). *Reference Dependence and Attribution Bias: Evidence from Real-Effort Experiments*. working paper.

Chen, Y., Harper, F. M., Konstan, J., & Li, S. X. (2010). Social comparisons and contributions to online communities: A field experiment on MovieLens. *The American Economic Review*, *100*(4), 1358-98.

Clark, A. E., Masclet, D., & Villeval, M. C. (2010). Effort and comparison income: Experimental and survey evidence. *ILR Review, 63(3),* 407-426.

Eyring, H., & Narayanan, V. G. (2018). Performance Effects of Setting a High Reference Point for Peer-Performance Comparison. *Journal of Accounting Research*, *56*(2), 581–615.

https://doi.org/10.1017/S1930297500008949 Published online by Cambridge University Press

Farber, H. S. (2008). Reference-dependent preferences and labor supply: The case of New York City taxi drivers. *The American Economic Review*, *98*(3), 1069–82.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics*, *108*(2), 437–459.

Gächter, S., & Thöni, C. (2010). Social comparison and performance: Experimental evidence on the fair wage–effort hypothesis. *Journal of Economic Behavior & Organization*, *76*(3), 531–543.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(3), 791-810.

Greiner, B., Ockenfels, A., & Werner, P. (2011). Wage transparency and performance: A real-effort experiment. *Economics Letters*, *111*(3), 236–238.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47(2),* 263–292.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, *121*(4), 1133-1165.

Kőszegi, B., & Rabin, M. (2007). Reference-dependent risk attitudes. *The American Economic Review*, *97*(4), 1047-1073.

Kőszegi, B., & Rabin, M. (2009). Reference-dependent consumption plans. *The American Economic Review*, *99*(3), 909–36.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442.

Rivas, M. F., & Sutter, M. (2008). *Wage dispersion and workers' effort* (No. 2008-15). Working Papers in Economics and Statistics.

Shalev, J. (2000). Loss aversion equilibrium. *International Journal of Game Theory*, *29*(2), 269–287.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.