Does the Perception of a Personalized, Algo-Driven Recommendation Causes More Trust in the Recommendation?

W241 Field Experiments Final Paper

David Tan

August 18, 2018

Introduction

It is increasingly common for businesses to provide personalized solutions to encourage more transactions and enhance customers satisfactions. Traditionally, sales staff may ask consumers certain questions regarding their background and preferences before providing a product or solutions that may suit the specific needs of the customer. In the age of big data and artificial intelligence, companies such as Netflix and Amazon pride themselves in having great algorithm to deliver very personalized and high quality recommendations. A study by Barilliance suggests the conversion rate of visitors who clicked on product recommendations was found to be 5.5 times higher than that of non clicking customers. Personalized "Top Sellers" recommendations has two times the click through rate of non personalized recommendations[1]. According to McKinsey, about 35% of Amazon's purchases and 75% of Netflix's video consumptions come from product recommendations in 2012[2]. Customers are clearly more motivated by recommendations that are personalized and/or algorithmic driven. The questions of interest are: how much of that motivation is due to actual improvement in recommendation quality? How much of it is due to simply the belief that the recommendation is personalized and algorithm driven?

Hypothesis and Potential Implications

To answer the question about whether or not customers are more motivated by recommendations that are personalized and/or algorithmic driven, we run a field experiment and hypothesize that the perception of a personalized algorithm driven recommendations would cause more trust in the recommendations.

If our hypothesis is supported by evidence from this research experiment, the findings can have widespread implications for any organizations or individuals who use persuasion to achieve their goals. If merely the perception of algo-driven recommendation alone can enhance persuasiveness, organizations should emphasize on enhancing the perception of the existence, or even the complexity and the intelligence of such algorithm. For consumers, the awareness of this psychological effect can help them be more cognizant when interpreting the quality of a recommendation that appeared to be personalized and algo-driven.

Furthermore, there could be a much bigger implications on socialcultural development. If the effect of having such perception is very significant, it may shed light on the magnitude of influence companies such as Google, Facebook, Amazon and Netflix can have in driving the collective development of culture through their recommendations.

Research Design

Our experiment draws upon two sets of control survey and two sets of investment survey with the following randomized design. Each respondent received two different surveys to increase the number of responses we can get given limitations on budget. Below we draw out using the ROXO grammer our experiment design.

Survey Set	Randomized Allocation	Control	Observations	Treatment	Observations
Survey Set 1	R	C_book	O	T_investment	O
Survey Set 2	R	C_investment	O	T_book	O

Both control and treatment survey ends with an identical recommendations on books or investment recommendations. Respondents are asked to rate the appropriateness of the recommendations and also the likelihood to consider and implement our recommendations.

The control surveys are very short with a few demographic questions. After the respondent rated the recommendations, we asked a subject specific question that we consider as a possible confounding variable. The treatment survey has a few more demographic questions (income, marital status, education) and about 10 subject specific questions (listed below). In addition, we explicitly stated that an algorithm is evaluating their inputs to generate a recommendations.

- Book Treatment Survey Subject Questions
  - On average, how many books do you read in a year? (confounding variable question in control survey post rating)
  - Do you currently own a book reading tablet (such as Kindle) for yourself?
  - Have you bought electronic books in the past?
  - Do you prefer reading off paper book or e-book?
  - Is learning about latest thoughts and ideas important to you?
  - Do you pick books based on some recommendation or best/selling list? If yes, select one or more.
  - On average, how long does it take for you to finish reading a book?
  - What is your most preferred method of purchasing books?

- Investment Treatment Survey Subject Questions
  - What is your household's current level of income?
  - Please rate the stability of your income
  - Do you have children?
  - How many dependents do you have
  - Is your living quarters owned or being bought by someone in the household, rented for cash, or occupied without payment of cash rent?
  - Estimate your annual living expenses and divide them by your available retirement fund
  - How strongly do you agree or disagree with the following statement? (Investment statement)
  - What are your objectives of financial planning?
  - Other than passive investment such as pension plans (e.g. 401k), are you experienced in asset allocation and active + + investment in stocks, bonds and/or derivatives? (confounding variable question in control survey post rating)
  - Investment knowledge question 1
  - Investment knowledge question 2
  - Investment knowledge question 3

Assumptions

We made a few key assumptions while operationalizing the hypothesis. We also incorporate checkpoints to evaluate if the assumptions might be violated.

1. Our survey setup can successfully create a perception that the recommendation is personalized and algorithmic driven.

We assume that a longer survey that collect more personal questions and explicitly state that the personal informations are being processed by an algorithm can successfully create the perception. To check whether the assumption is violated, treatment respondents are asked to rate whether they believe the recommendations are generated by an algorithm after they rate the recommendations.

2. Survey respondents fill out the survey diligently.

We incorporated attention check mathematical questions to filter out respondents who cruise through the survey with random clicks. We also incorporated open ended questions such as "How do you feel today?" to ensure engagement.

3. There is no spillover from control survey to treatment survey.

To prevent respondents' perception that the answers they provide in the treatment survey would affect the recommendation in the control survey, every respondent receive control survey and recommendation before receiving treatment survey and recommendation. In addition, prior to receiving treatment survey, we explicitly stated that the two surveys are completely different, and asked respondents whether the two surveys are independent.

4. Our survey questions and recommendations are appropriate

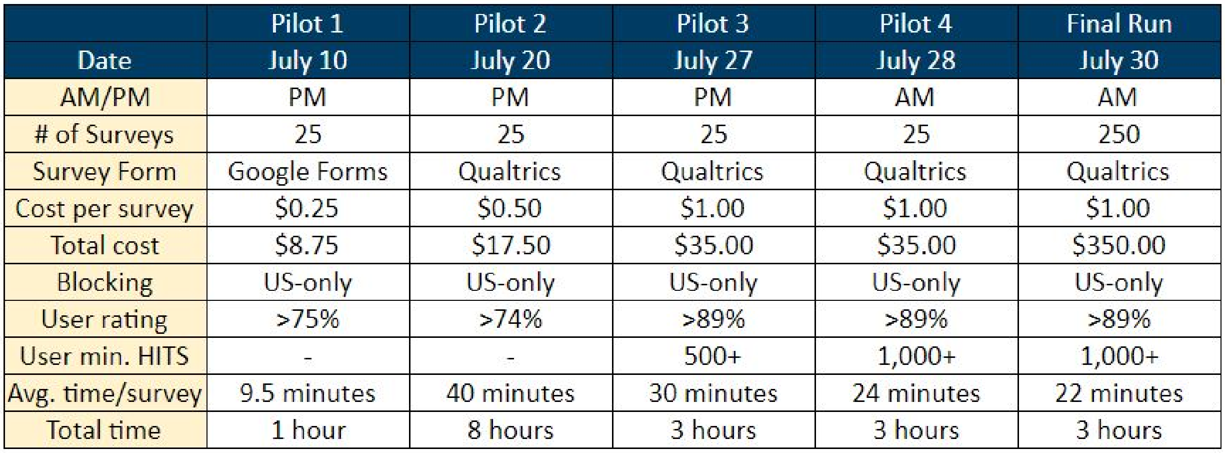
We chose recommendations that are very generic and most likely neutral to an average population. We selected the three self-help books from Amazon and New York Times best selling lists and the selection was based on the belief that most people would find them neutral and inoffensive. The investment advice of 35.8% Equity, 27.7% Fixed Income, 36.5% Alternative Investment is a slight deviation of a general advice of having about 60% in equity and 40% bond for a passive investor (also known as the classic 60-40 portfolio). We are cognizant that our respondents from Mechanical Turks tend to be younger and thus the advice is less likely to be perceived as very inappropriate for a retired person who would more likely find lower risk portfolio more appropriate.

While the subject specific questions are closely related to the recommendations, we ensured that no answer from the survey would contradict the recommendations. This is to prevent respondents from having a clue that the recommendation is independent of their inputs.

Experiment Execution

To assess and determine the appropriate format and survey questions, we interviewed representatives from Amazon Mechanical Turk ("mTurk") and Qualtrics to determine the most feasible option given budget constraints and to learn best practices to maximize survey effectiveness. mTurk was chosen over Qualtrics for recruiting respondents due to flexibility in pricing. We conducted the first pilot on Google Forms initially, but switched to Qualtrics in the second pilot as Qualtrics provided more useful functions such as random assignment of surveys and better design features. We also incorporated suggestions from Prof. Alex Hughes and Prof. David Reiley, such as utilizing certain features on Qualtrics (such as randomly generated completion code) and applying blocking to include only US-based subjects (to prevent respondents being entirely from a developing country).

There were four rounds of pilot studies, each with different variations highlighted in figure 1. Other than changing the survey platform from Google Forms to Qualtrics, we tested different reward amounts, time limitations, user ratings, and the time of the day in which the pilots were launched. We also inserted a few attention checks in the process and adjusted the survey based on users' feedback and our observations.



	Pilot 1	Pilot 2	Pilot 3	Pilot 4	Final Run
Date	July 10	July 20	July 27	July 28	July 30
AM/PM	PM	PM	PM	AM	AM
# of Surveys	25	25	25	25	250
Survey Form	Google Forms	Qualtrics	Qualtrics	Qualtrics	Qualtrics
Cost per survey	$0.25	$0.50	$1.00	$1.00	$1.00
Total cost	$8.75	$17.50	$35.00	$35.00	$350.00
Blocking	US-only	US-only	US-only	US-only	US-only
User rating	>75%	>74%	>89%	>89%	>89%
User min. HITS	-	-	500+	1,000+	1,000+
Avg. time/survey	9.5 minutes	40 minutes	30 minutes	24 minutes	22 minutes
Total time	1 hour	8 hours	3 hours	3 hours	3 hours

Figure 1: Pilot Experiments Timeline

The quality of the responses reached our satisfaction by the 4th pilots. We observed that the quality of responses has improved with the inclusion of attention checks, time limitations and increasing mTurk user rating requirement. Even though there were only 25 respondents in the 4th pilot, the results were by and large similar to the results from the actual survey. The final iteration of our survey design provided the mTurk user with a link to the Qualtrics website, where they would be randomly assigned to either Set 1 or Set 2. Whichever set the subject was assigned to, all subjects would respond to 2 surveys: a control survey prior to a treatment survey to prevent spillover.

Survey Mechanism

A subject assigned to set 1 of the surveys would receive book recommendations (control) first and portfolio investment recommendation (treatment) next. A subject assigned to set 2 on the other hand would receive portfolio investment recommendation (control) first and book recommendations (treatment) next. Figure 2 below shows our survey structure. Depending on whether the subject answered 1 designated attention check quiz question correctly, the subject would see a unique completion code at the end to enter into both the Qualtrics survey and into mTurk. The code was a six-digit number that was randomly generated by Qualtrics. This allowed us to match the survey answers to the mTurk subjects and address any issues.

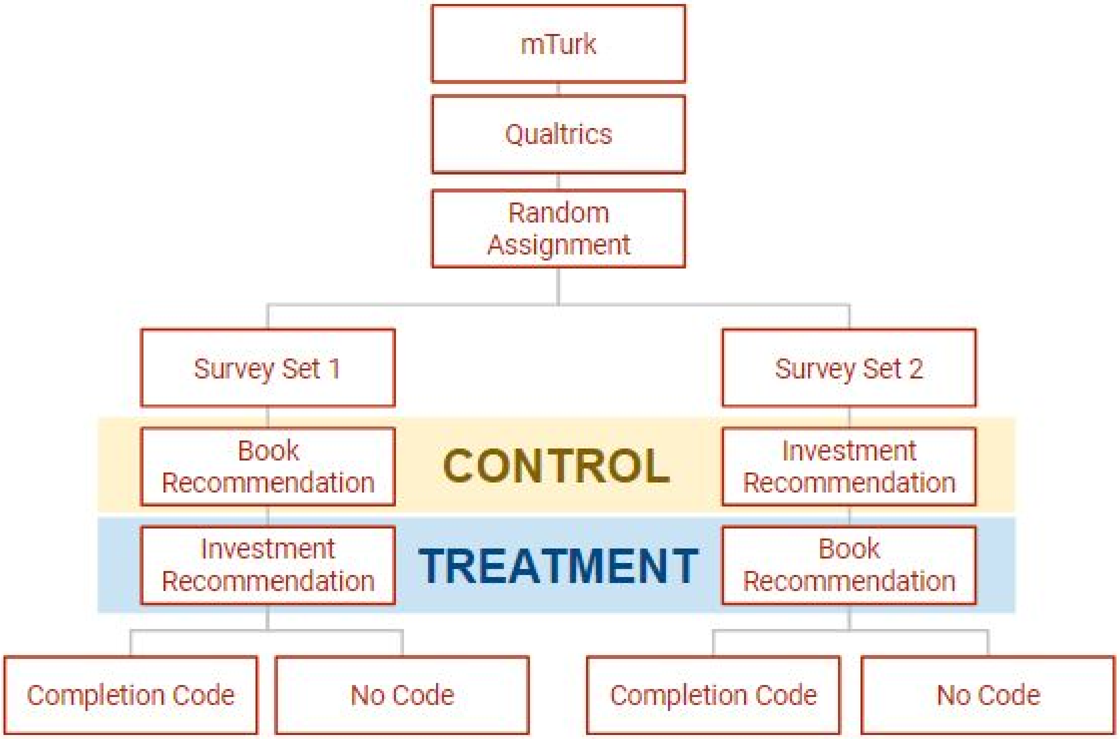
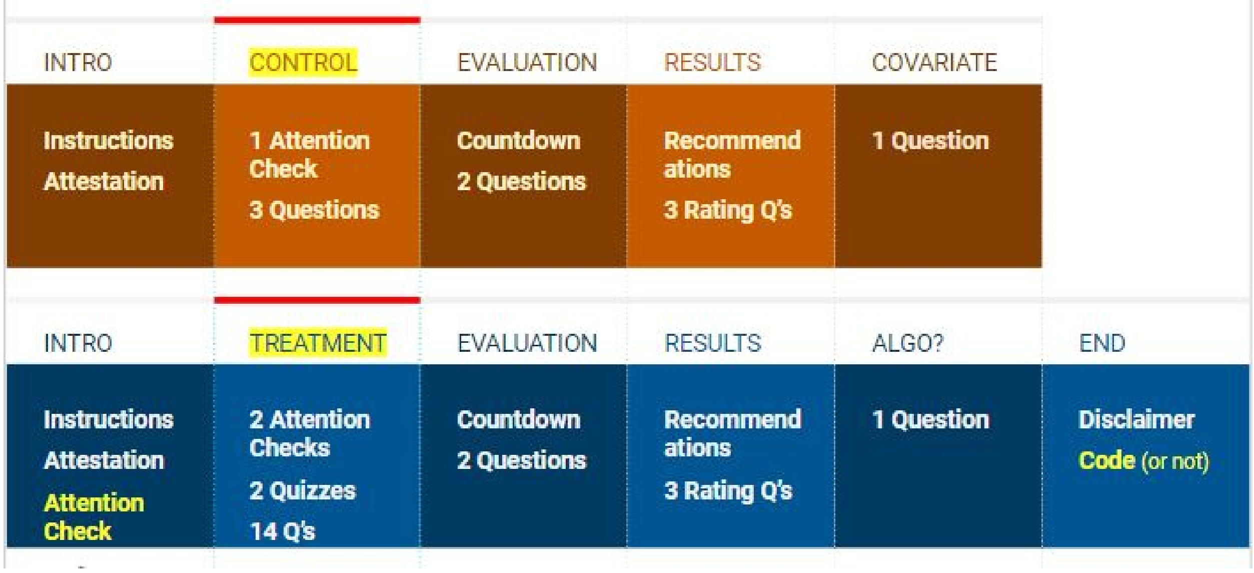


Figure 2: Survey Structure

After several iterations throughout our pilot experiments, our survey was set to contain 6 attention checks. Three of them were arithmetic quizzes deliberately written out in sentences. Three of them were long answer questions that were implemented to engage the users. This quiz question was placed between the first and second surveys. After explicitly explaining that the second survey is not related to the first survey, we asked the users if the second survey is affected by the answers in the first survey. If the user correctly answered "No", they would eventually see the completion code upon completion of the surveys. If they answered "Yes" or "Maybe", they would see a message explaining why they do not see the code.

The attention checks allowed us to measure the quality of survey responses. The quiz questions along with long answer questions were useful in assessing whether or not the respondent was paying attention to the questions and answer choices. In the introduction, respondents are told that their answers are not shared outside our research and that we will not try to personally identify the users with the answers they provide. They were also warned to watch out for the attention checks as they may affect their chances of getting the completion code at the end.

The control part of the survey contains 4 questions, one of which was an attention check. Three simple demographic questions were followed with a 10-second countdown to give the impression that an algorithm may be running. The page was followed by recommendations that had no influence from the answers given by the users. For the group that went through the first set, 3 books were recommended. The group that went through the second set would see a portfolio asset allocation recommendation that gave percentage recommendations for equity, fixed income, and alternative investments. At the bottom of the recommendation, the user would rate the appropriateness of the recommendations, and the likelihood of implementation of the recommendations. One more question was asked afterwards to measure the number of books a user reads in a year, or experience in investments.



INTRO	CONTROL	EVALUATION	RESULTS	COVARIATE
Instructions Attestation	1 Attention Check 3 Questions	Countdown 2 Questions	Recommendations 3 Rating Q’s	1 Question

INTRO	TREATMENT	EVALUATION	RESULTS	ALGO?	END
Instructions Attestation Attention Check	2 Attention Checks 2 Quizzes 14 Q’s	Countdown 2 Questions	Recommendations 3 Rating Q’s	1 Question	Disclaimer Code (or not)

Figure 3: Survey Flowchart

The second survey, the treatment, would start similarly as the control. After instructions and attestation, the user would be asked to answer the attention check quiz question that would be decisive in generating the completion code at the end. The treatment survey consisted of 14 questions regarding the user or the user’s preferences, 2 attention check questions, and 2 quiz questions. It is a much longer survey compared to the control part of the survey.

Similar to the control part, the user would see a 10 second countdown, followed by recommendations and rating questions. A final question would be asked on whether the user believed the recommendations were generated by an algorithm that took the users’ responses as input. A comparison in the number of questions between control part and treatment part is shown in the flow diagram above in figure 3.

At the end of the survey, the users were provided a disclaimer explaining that this was an academic experiment and that the recommendations should not be considered seriously.

Data Analysis

Data Cleanup

We have received 322 responses in total. Among them, we removed 56 (17.4%) subjects who did not finish the survey. As we mentioned earlier, we included attention checks for each respondent to control data quality. There are 60 (18.6%) subjects who gave wrong answers for all arithmetic attention check questions, thus we had to remove these entries as poor quality responses. Overall, we’ve removed 116 (36%) subjects from the raw survey data due to incompleteness and bad data quality. After loading in the raw data, we assign numeric values to the character responses. Our model covariates are coded as below:

- Independent Variables:
  - **Age group** (18-24 = 1, 25-34 = 2, 35-44 = 3, 45-54 = 4, 55+ = 5)
  - **Gender** (Female = 0, Male = 1)
  - **Income** (Less than \$20k = 1, \$20k-\$35k = 2, \$35k-\$50k = 3, \$50k-\$75k = 4, \$75k-\$100k = 5, \$100k+ = 6)
  - **Education** (below bachelor = 0, bachelor or above = 1)
  - **Marital Status** (unmarried = 0, married = 1)
  - **Average books read per year**

  - **Years of investment experience**
- Dependent Variables:
  - **Book survey recommendation rating** (Inappropriate = 1, Neutral = 2, Appropriate = 3)
  - **Book survey recommendation consideration** likelihood (Unlikely=1, Neutral=2, Likely = 3)
  - **Investment (invm) survey recommendation rating** (Inappropriate = 1, Neutral = 2, Appropriate = 3)
  - **Invm survey recommendation consideration likelihood** (Unlikely=1, Neutral=2, Likely = 3)

We separate control and treatment groups from the book and investment survey data. For the book survey, we have 114 subjects in control and 92 subjects in treatment. For the investment survey, we have 92 subjects in control and 114 subjects in treatment. We compare the treatment effect of book survey and investment survey separately as we believe that the inherent level of appropriateness of our book and investment recommendations might be different and thus would be invalid as an apple to apple comparison. We first compute the average treatment effect (ATE) for the book survey.

Average Treatment Effect

```
##                                          ATE
## book recommendation           -0.135392830
## book consideration likelihood -0.003432494
```

The estimated ATE for book recommendation rating is -0.1354 and book consideration likelihood is -0.003. Note that both ATE values are negative, which indicates that treatment group subjects provide lower recommendation rating and lower likelihood than the control group.

Below are the percentages of subjects in each dependent variable's categories between control and treatment.

```
##                               Inappropriate % Neutral % Appropriate %
## control book recommendation               9.6      23.7          66.7
## treatment book recommendation            15.2      26.1          58.7
```

```
##                                         Unlikely % Neutral % Likely %
## control book consideration likelihood         20.2      20.2     59.6
## treatment book consideration likelihood       20.7      19.6     59.8
```

66.7% of control group subjects chose "Appropriate" for book recommendation, while 58.7% in treatment group chose "Appropriate". 9.6% of control group subjects chose "Inappropriate" for book recommendation, compared to 15.2% of treatment group chose "Inappropriate".

59.6% of control group subjects chose "Likely" for book consideration likelihood, while 59.8% in treatment group chose "Appropriate". 20.2% of control group subjects chose "Unlikely" for book consideration likelihood, compared to 20.7% of treatment group chose "Unlikely".

Next, we compute the frequency table and compute the average treatment effect of investiment survey.

```
##                                        ATE
## invm recommendation             0.01773455
## invm consideration likelihood -0.04271548
```

The estimated ATE for investment recommendation is 0.0177 and investment consideration likelihood is -0.0427. Note that ATE for investment recommendation is positive while the ATE for investment consideration likelihood is negative. These numbers send us mixed signals and we need to conduct other tests to fully understand the data.

```
##                               Inappropriate % Neutral % Appropriate %
## control invm recommendation               9.8      53.3          37.0
## treatment invm recommendation            8.8      53.5          37.7
```

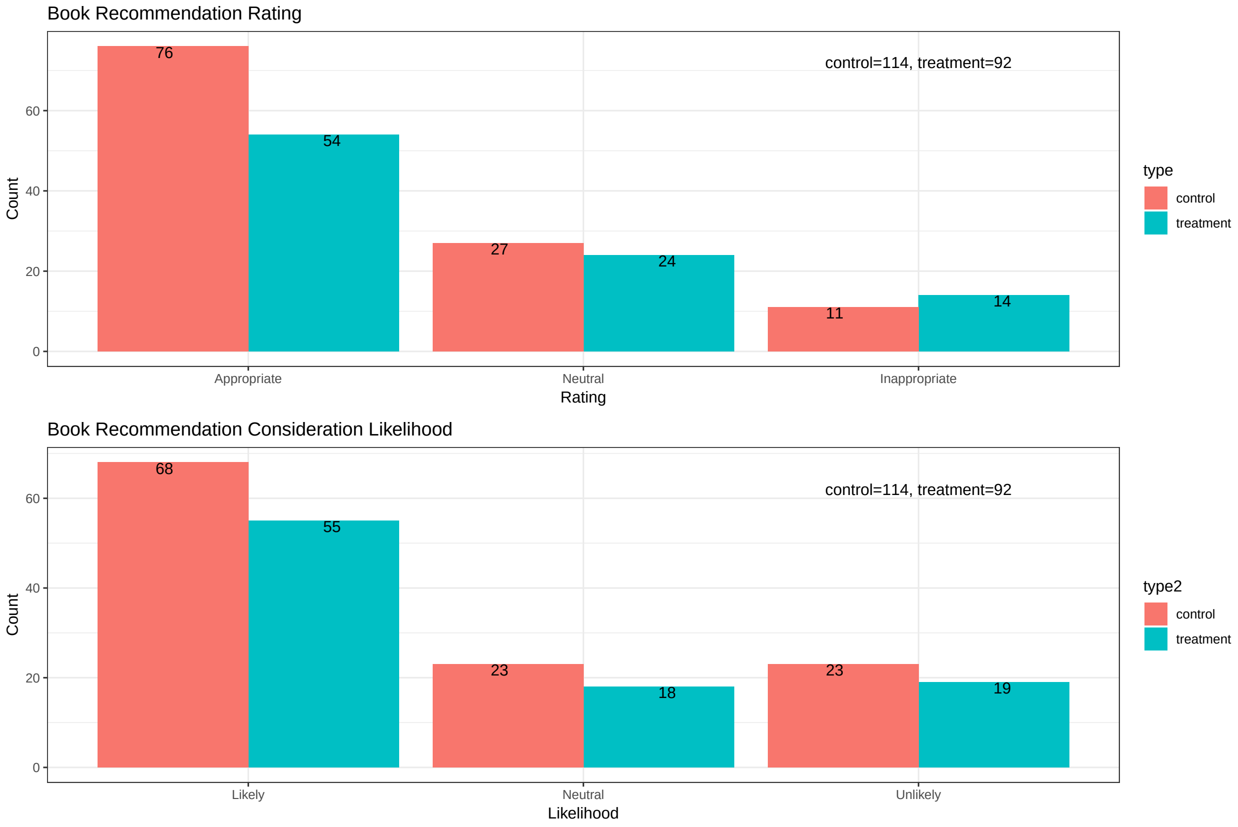
```
##                                    Unlikely % Neutral % Likely %
## control invm consider likelihood         27.2      32.6     40.2
## treatment invm consider likelihood       22.8      45.6     31.6
```

37.0% of control group subjects chose "Appropriate" for investment recommendation, while 37.7% in treatment group chose "Appropriate". 9.8% of control group subjects chose "Inappropriate" for book recommendation, compared to 8.8% of treatment group chose "Inappropriate". Note that the "neutral" samples are relatively large.

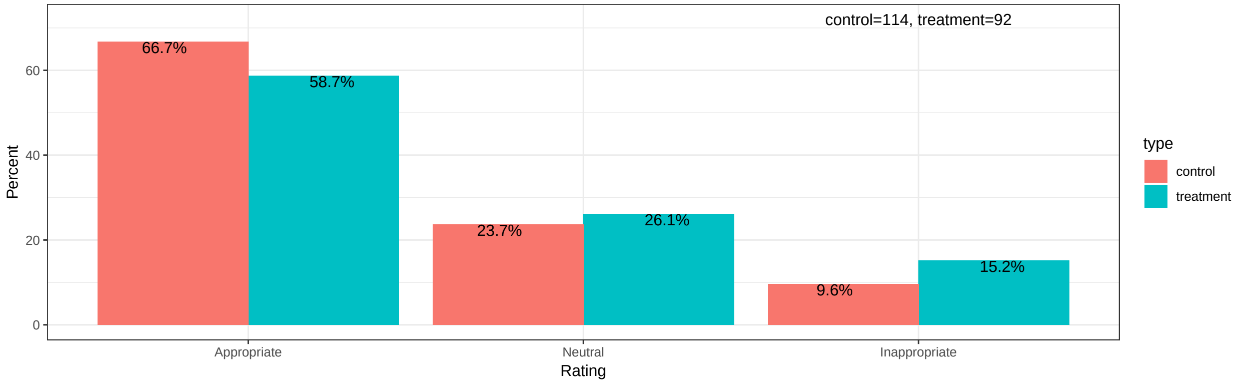
40.2% of control group subjects chose "Likely" for investment consideration likelihood, while 31.6% in treatment group chose "Likely". 27.2% of control group subjects chose "Unlikely" for invm consideration likelihood, compared to 22.8% of treatment group chose "Unlikely".

Outcome measures

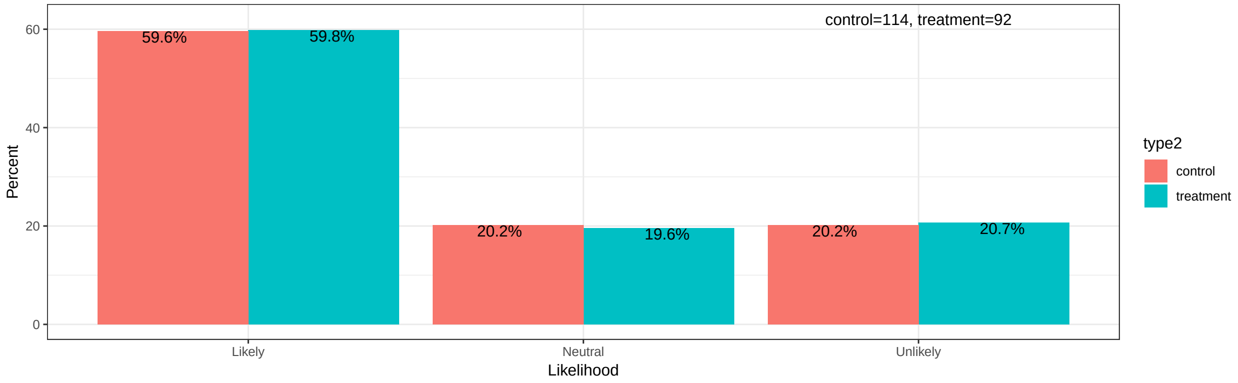
Recall that we have two response outcomes for each survey: recommendation rating and recommendation consideration likelihood. They are ordered categorical variables with three levels. Below are the outcomes distribution charts.They look normal as we don't see any unexpected numbers.



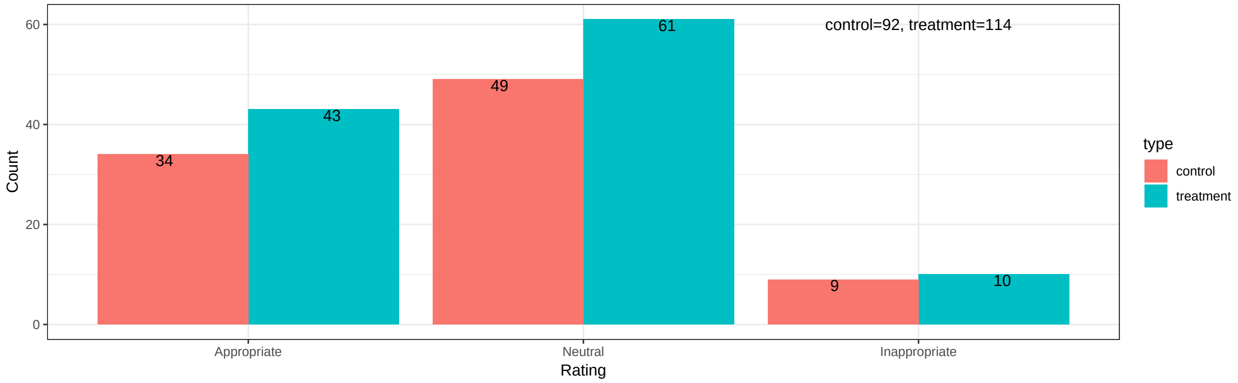
Book Recommendation Rating



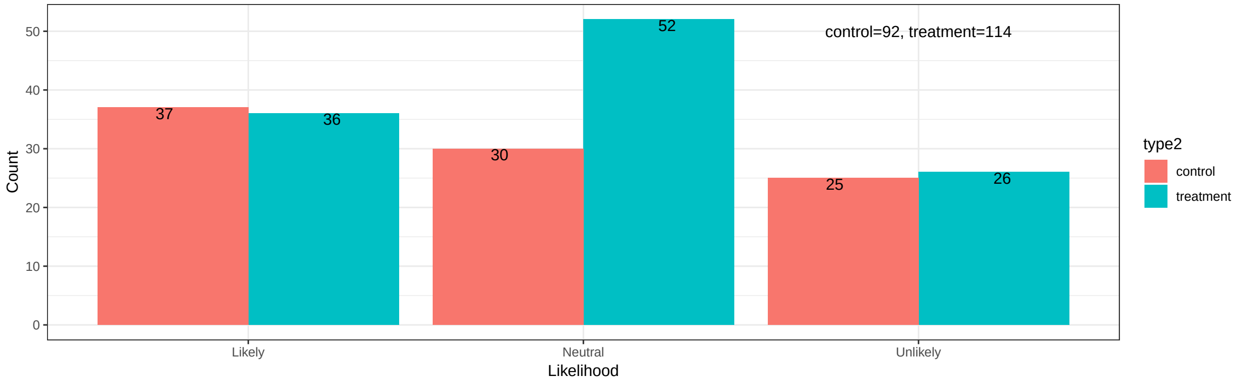
Book Recommendation Consideration Likelihood

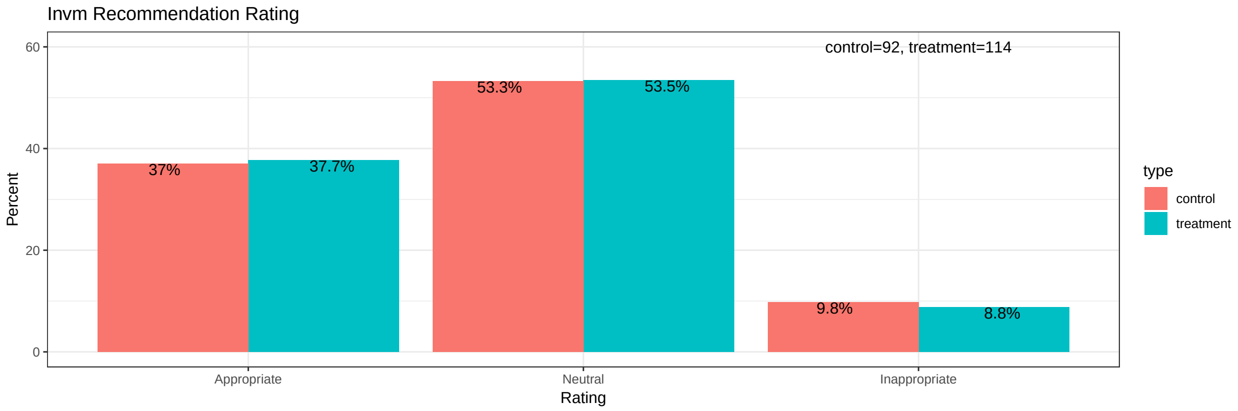


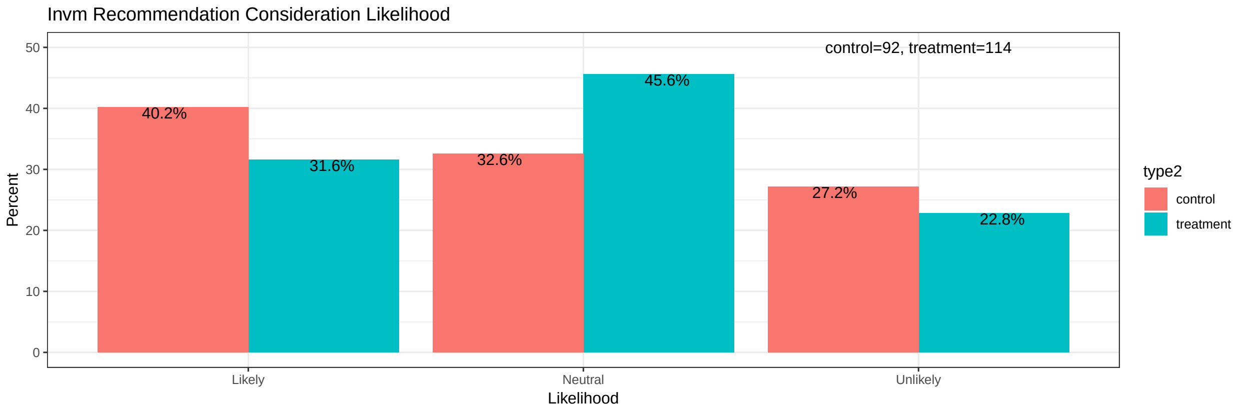
Invm Recommendation Rating



Invm Recommendation Consideration Likelihood







Statistical power

To collect a broad sample of the American population, we turned to Mechanical Turk to select our subjects. We used Qualtrics to randomly assign subjects to either survey set 1 or set 2. Prior to commencing the analysis of the experiment, we wanted to know whether or not there is a chance of detecting an effect. To do that, the average recommendation rating and the standard deviation of recommendation rating must be identified for the control population.

Power analysis is an important aspect of experimental design. It allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence. Conversely, it allows us to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints.

```
##
##      t test power calculation
##
##              n1 = 114
##              n2 = 92
##               d = 0.2036651
##       sig.level = 0.05
##           power = 0.3040989
##     alternative = two.sided
```

```
##
##      t test power calculation
##
##              n1 = 92
##              n2 = 114
##                 d = 0.02811561
##         sig.level = 0.05
##             power = 0.05457954
##       alternative = two.sided
```

Book survey statistical power was calculated to be 0.3041, based on the book recomendation rating. It means that a survey (when conducted repeatedly over time) is likely to produce a statistically significant results 30.4 times out of 100. In other words, we only has 30.4% chance of finding a statistically significant treatment effect when such an effect really exists.

Investment survey statistical power was calculated to be 0.0546, based on the investment portfolio recommendation rating. It means that a survey (when conducted repeatedly over time) is likely to produce a statistically significant results 5.5 times out of 100. In other words, we only has 5.5% chance of finding a statistically significant treatment effect when such an effect really exists.

Statistical power = 1 - $P[TypeIIerror]$ = probability of finding an effect that is there. Given the low values of statistical power. We will likely to find a false negative when there is an effect in the population, but we fail to reject the null hypothesis (there is no effect).

Assume that the control and treatment have similar variance, if we want 80% power in detecting a rating score difference between the control and treatment groups using $\alpha = 0.05$, the groups would need to be properly sized.

```
##
##      Two-sample t test power calculation
##
##               n = 379.4092
##           delta = 0.1353928
##              sd = 0.6647815
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

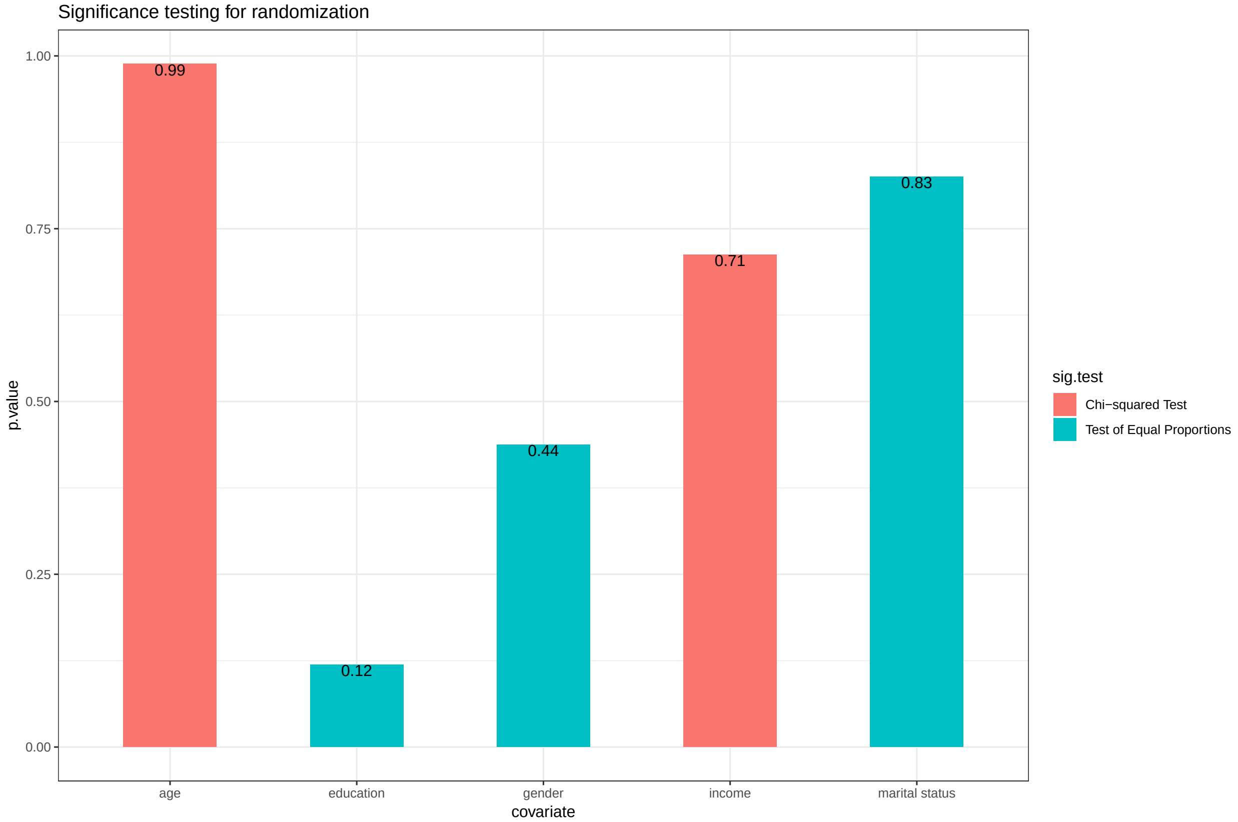
We would need to have at least 380 subjects in each control and treatment group to have a statistically significant results 80% of the times.

Verify randomization

This step is to confirm that the assignment to treatment and control groups is random. It must be verified that the proportion of individuals assigned to each group was similar, thus providing a confidence level that there is internal validity in the randomization.

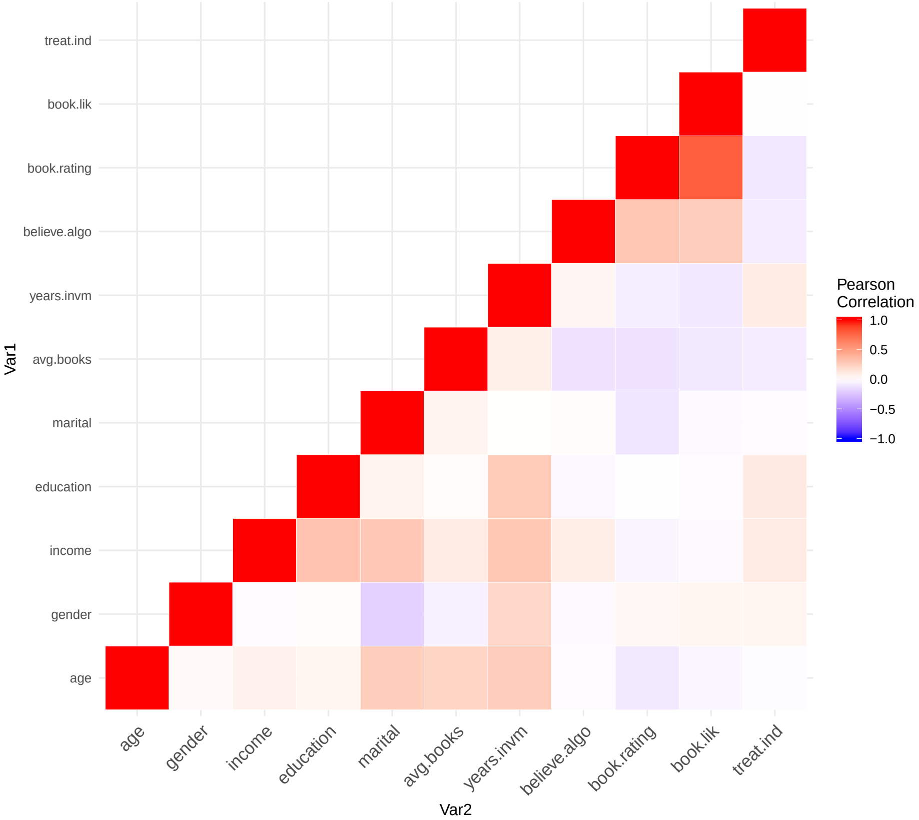
Two-sample test for equality of proportions (`prop.test`) and Pearson's Chi-squared test (`chisq.test`) were used to determine if each demographic covariate's percentages are well balanced between control and treatment group. The p-values for each test are listed below. Note that the control subjects of the book survey is also the treatment subjects of the investment survey and the treatment subjects of the book survey is also the control subjects of the investment survey, we can only run tests once based on one survey results. Since the p-values are all much greater than 0.05, we have insufficient evidence to conclude that the distributions of each covariate is different for control and treatment groups. This gives us confidence that our randomization was correctly implemented.

```
##        covariate p.value                  sig.test
## 1         gender  0.4375 Test of Equal Proportions
## 2            age  0.9885           Chi-squared Test
## 3 marital status  0.8254 Test of Equal Proportions
## 4      education  0.1191 Test of Equal Proportions
## 5         income  0.7124          Chi-squared Test
```



Correlation

The next step is to look at the covariates that were collected in the questionnaire. As they were unrelated to treatment, we expect that there will be no correlation within covariates. However, we expects that these covariates will explain some of the variance in the outcome.



Based on the correlaion matrix headmap, there is no evidence of multi-colinearity among covariates. In addition, the subjects of book survey are also the subjects of investment survey. So the correlation results for the two survyes are the same.

Ordinal Logistic Regression

We now turn our attention to models for our outcome measures. Recall that the outcomes are

1. "recommendation rating (Inappropriate = 1, Neutral = 2, Appropriate = 3)" and
2. "recommendation consideration likelihood (Unlikely=1, Neutral=2, Likely = 3)" and they are ordered categorical variables.

Thus, ordered logit models can be fitted in R using the `polr` function, short for proportional odds logistic regression, in the package `MASS`. For each survey, We will treat recommendation rating and recommendation consideration likelihood as the two responses, and we will fit models for each response. Demographic covariates and other attributes are model predictors.

For the book survey, we try to fit the following four models:

- baseline:

$$rating_{book} = \beta_0 + \beta_1 treat.ind$$

- model.demo:

$$rating_{book} = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 income + \beta_4 education + \beta_5 marital + \beta_6 treat.ind$$

- model.attr:

$$rating_{book} = \beta_0 + \beta_1 avg.book + \beta_2 treat.ind$$

- model.full:

  $rating_{book} = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 income + \beta_4 education + \beta_5 marital + \beta_6 avg.book + \beta_7 treat.ind$

Table 2: Book Recommendation Models Comparison

	Dependent variable:			
	Book Recommendation Rating			
	(1)	(2)	(3)	(4)
age		−0.128		−0.079
		(0.142)		(0.147)
gender		0.069		0.027
		(0.292)		(0.295)
income		0.014		0.029
		(0.105)		(0.107)
education		0.042		0.052
		(0.305)		(0.306)
marital		−0.280		−0.304
		(0.312)		(0.313)
avg.books			−0.210	−0.192
			(0.129)	(0.135)
treat.ind	−0.371	−0.398	−0.418	−0.444
	(0.284)	(0.289)	(0.286)	(0.292)
Observations	206	206	206	206
Note:		*p<0.1; **p<0.05; ***p<0.01		

We also fit models for the second outcome: "Book recommendation consideration likelihood".

Table 3: Book Consideration Likelihood Models Comparison

	Dependent variable:			
	Book Recommendation Consideration Likelihood			
	(1)	(2)	(3)	(4)
age		−0.044 (0.136)		−0.016 (0.140)
gender		0.118 (0.285)		0.106 (0.285)
income		0.005 (0.100)		0.018 (0.102)
education		−0.088 (0.293)		−0.087 (0.293)
marital		−0.040 (0.304)		−0.045 (0.304)
avg.books			−0.128 (0.125)	−0.123 (0.129)
treat.ind	−0.003 (0.277)	−0.004 (0.282)	−0.030 (0.278)	−0.033 (0.283)
Observations	206	206	206	206
Note:			*p<0.1; **p<0.05; ***p<0.01	

For investment survey, we also list our model equations.

- baseline:

$$rating_{invm} = \beta_0 + \beta_1 treat.ind$$

- model.demo:

$$rating_{invm} = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 income + \beta_4 education + \beta_5 marital + \beta_6 treat.ind$$

- model.attr:

$$rating_{invm} = \beta_0 + \beta_1 years.invm + \beta_2 treat.ind$$

- model.full:

$$rating_{invm} = \beta_0+\beta_1 age+\beta_2 gender+\beta_3 income+\beta_4 education+\beta_5 marital+\beta_6 years.invm+\beta_7 treat.ind$$

Table 4: Invm recommendation rating

	Dependent variable:			
	Investment Recommendation Rating			
	(1)	(2)	(3)	(4)
age		0.042		0.066
		(0.138)		(0.144)
gender		0.151		0.187
		(0.280)		(0.286)
income		−0.012		0.003
		(0.101)		(0.104)
education		0.020		0.045
		(0.284)		(0.287)
marital		0.299		0.282
		(0.305)		(0.306)
years.invm			−0.039	−0.075
			(0.109)	(0.122)
treat.ind	0.050	0.046	0.041	0.040
	(0.272)	(0.276)	(0.273)	(0.277)
Observations	206	206	206	206
Note:	*p<0.1; **p<0.05; ***p<0.01			

We also fit models for the second outcome: "investment recommendation consideration likelihood".

Model results are shown in Table 2-5. We check the coefficient estimates (estimated ATE) and the p-value results but no covariate is statistically significant in all model fittings. Therefore we don't have evidence to reject the null hypothesis that there is no treatment effect.

Table 5: Invm recommendation consideration likelihood

	Dependent variable:			
	Investment Recommendation Consideration Likelihood			
	(1)	(2)	(3)	(4)
age		−0.173		−0.152
		(0.135)		(0.138)
gender		0.043		0.073
		(0.265)		(0.269)
income		−0.089		−0.074
		(0.098)		(0.101)
education		−0.224		−0.191
		(0.274)		(0.279)
marital		0.048		0.025
		(0.286)		(0.287)
years.invm			−0.141	−0.079
			(0.105)	(0.115)
treat.ind	−0.127	−0.189	−0.176	−0.204
	(0.261)	(0.266)	(0.264)	(0.267)
Observations	206	206	206	206
Note:				*p<0.1; **p<0.05; ***p<0.01

Model Discussion

Our study did not find any statistically significant causal effect of the perception of an algo-driven recommendation on more trust in recommendations. While it is possible that this causal relationship do not exist, we argue that our failure to find a significant effect is affected by the survey's failure to cause all respondents to have such perception, as evidenced by the respondents' feedback on whether they believe the recommendation is generated by an algorithm (1 - No, 2 - Maybe, 3 - Yes).

```
##                1  2  3
## T_book        35 30 27
## T_investment 36 38 40
```

We only successfully convinced about a third of the respondents to fully believe that the recommendation is backed by an algorithm. Such response shows that the first assumption - Our survey setup can successfully create a perception that the recommendation is personalized and algorithmic driven - was violated and that our experimental setup is not ideal to test our hypothesis.

There is also evidence that our recommendations may not be general enough. Some respondents commented that they "hate self-help books", "do not understand investment enough to know if the recommendation is appropriate", "do not understand what is alternative investment". This shed light on the difficulty in operationalizing our hypothesis. However neutral and general a recommendation is, there is very few recommendations that would work for every person in the world. When our sample size is relatively small with about 90-100 in each treatment, the responses could be skewed by a disproportionate number of people who feel very strongly about something we perceive as relatively neutral (self-help books) to a general population. This problem is ideally neutralized by randomization, but it would work better if we have larger sample size.

While our experiments failed to find statistically significance evidence to support our hypothesis, we noticed positive correlations between a respondents' ratings and his/her belief in whether the recommendations is generated by an algorithm. As both observations are collected after treatment, we cannot conclude that stronger belief caused a higher rating. The finding is certainly interesting as it suggests positive association betweein recommendations quality and algo-driven recommendations. We do not know whether it is the perception (believe that there is an algorithm) that causes the higher ratings, or whether it is the appropriateness that causes the respondents to believe that there is an algorithm. It is also possible that these subjects are more impressionable in general and believe in all our claims in the survey - both the recommendation and our statement that there is a computer algorithm processing their inputs.

Correlation between rating and belief in the existence of an algo for book and investment treatment:

```
## 0.4754 , 0.2934
```

Although it was not the original intent of our study, we investigate whether certain types of people are more easily convinced by our design setup. We regressed "believe in algo" based on three different datasets. First is the only book survey treatment group. Second is the investment group. And third is all treatment group.

- model.book:

$$believe.algo = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 income + \beta_4 education + \beta_5 marital + \beta_6 avg.books$$

- model.invm:

$$believe.algo = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 income + \beta_4 education + \beta_5 marital + \beta_6 years.invm$$

- model.att:

$$likelihood_{invm} = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 income + \beta_4 education + \beta_5 marital + \beta_6 years.invm + \beta_7 avg.books$$

In table 6, notice that the coefficient estimate (estimated ATE) of "avg.books" is statistically significant at 90% confidence level. However the coefficient estimate is negative, which indicates that the more books people read the more skeptical they are. The belief in the existence of an algorithm is otherwise not affected by age, gender, income, or education.

Table 6: Believe in Algo

	Dependent variable:		
	Book, Investment, All.Treatment		
	(1)	(2)	(3)
age	−0.048	−0.056	0.010
	(0.210)	(0.187)	(0.139)
gender	−0.355	0.241	−0.146
	(0.408)	(0.365)	(0.269)
income	0.232	0.028	0.135
	(0.146)	(0.132)	(0.098)
education	0.108	−0.503	−0.276
	(0.421)	(0.364)	(0.275)
marital	−0.383	0.153	−0.066
	(0.425)	(0.391)	(0.288)
avg.books	−0.066		-0.229^{*}
	(0.173)		(0.119)
years.invm		0.116	0.087
		(0.163)	(0.115)
Observations	92	114	206

Note: *p<0.1; **p<0.05; ***p<0.01

Conclusions

Our experiments could not find statistically significant treatment effect to support our hypothesis that the perception of a personalized algorithmic driven recommendation causes more trust in the recommendations. Such result is at least partly caused by the inability of our design to implant such perception in the respondents' mind. This highlights the difficulty of evaluating treatment effect when the treatment is to create a psychological impression. While our studies could not prove causual effect, it illustrated positive correlation between the perception and trust in the recommendations, and that the perception is negatively associated with how much a respondent read. To further improve our experiment, specifically the operationalization of our hypothesis, we should keep modifying our survey design until we are convinced that the design is effective in creating the desired perception in the respondents' minds.

Reference

[1] Personalized Product Recommmendation Stats. (n.d.) Retrieved Jun 6, 2018, from https://www.barilliance.com/personalized-product-recommendations-stats/

[2] How retailers can keep up with consumers. (2013, Oct) McKinsey & Company. Retrieved Jun 6, 2018 from https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

[3] Power Analysis, https://www.statmethods.net/stats/power.html

[4] It's the Effect Size, Stupid What effect size is and why it is important https://www.leeds.ac.uk/educol/documents/00002182.htm

[5] Getting started with the pwr package https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html