BANK OF CANADA
BANQUE DU CANADA

Staff working paper / Document de travail du personnel—2026-5

Last updated: March 4, 2026

I Am So Tired! I Don’t Know What to Do! Survey Fatigue and Financial Literacy: Results from a Randomized Experiment

Anna Chernesky
annachernesky@gmail.com

Marcel C. Voia
Laboratoire d’Économie d’Orléans, Université d’Orléans, University of Bucharest, and Department Special Adviser, Bank of Canada
marcel.voia@univ-orleans.fr

Kim P. Huynh
Indiana University Bloomington and Laboratoire d’Économie d’Orléans, Université d’Orléans
kphuynh@iu.edu

Bank of Canada staff research is produced independently from the Bank’s Governing Council and may support or challenge prevailing views. The views expressed in this paper are solely those of the authors and may differ from official Bank of Canada positions. No responsibility for them should be attributed to the Bank.

DOI: https://doi.org/10.34989/swp-2026-5 | ISSN 1701-9397

Acknowledgements

We thank Badi Baltagi, Heng Chen, Theodoros Garanzotis, Derek Messacar, Gary Mottola, and Angelika Welte as well as the participants of the Conference in Memorial of Anand Acharya and Charles Saunders at Carleton University (9 November 2024), Babeş–Bolyai University, Clermont-Ferrand, Memorial University, 2025 Current Innovations in Probability-Based Household Internet Panel Research (CIPHER) Conference, and the Laboratoire d'Économie d'Orléans for their constructive comments. We thank Diana Crossan, David Fernandez, and Annamaria Lusardi for encouraging this research. We thank Marie-Hélène Felt for her assistance in the nascent stages of this project. We thank Maren Hansen and Meredith Fraser-Ohman for editorial assistance. This work was completed while Anna Chernesky was a research assistant at the Bank of Canada. Kim Huynh and Marcel Voia dedicate this paper to the memory of Anand Acharya and Charles Saunders, who encouraged us to be lifelong learners. Marcel Voia thanks the Bank of Canada, Currency Department for the support of this project. The title of this paper is inspired by the Beatles' song "I'm so tired" from their 1968 double album "The Beatles" (also known as "the White Album"). The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada.

Abstract

Cross-country evidence finds that there are low levels of financial literacy. Financial literacy is often measured using the “Big Three” questions about interest rates, inflation, and risk. These questions are usually part of a longer survey. Respondents in long surveys may suffer survey fatigue and have lower quality responses. Therefore, the placement of the questions (and survey fatigue) may play a role in the results. We use a randomization of question placement to estimate the causal effect on financial literacy results. We find that when financial literacy questions are placed at the end of a survey, respondents are more likely to answer “Don’t know.” The increase in “Don’t know” responses comes largely at the expense of correct responses. We find that this leads to a drop in financial literacy by 5%-15%. This research suggests a measure of financial literacy that is adapted to account for survey length.

Research themes: Econometric, statistical and computational methods; Cash and bank notes; Digital assets and fintech; Payment and financial market infrastructures; Retail payments
JEL codes: C81, C83, D12, G53

Résumé

Des données comparatives entre pays montrent que les niveaux de littératie financière sont faibles. La littératie financière est souvent mesurée à l’aide de trois questions principales portant sur les taux d’intérêt, l’inflation et le risque. Ces questions font généralement partie d’une enquête plus longue. Les personnes qui participent à de longues enquêtes peuvent ressentir de la lassitude et fournir des réponses de moindre qualité. Ainsi, la position des questions (et la lassitude liée à l’enquête) peut influencer les résultats. Nous utilisons une répartition aléatoire des questions pour estimer l’effet causal de leur positionnement sur les résultats de littératie financière. Nous observons que lorsque les questions mesurant la littératie financière sont placées à la fin d’une enquête, les personnes interrogées sont plus susceptibles de répondre « Je ne sais pas ». Cette augmentation se fait en grande partie au détriment de réponses correctes. Nous constatons que cela entraîne une baisse de 5 % à 15 % des résultats de littératie financière. Cette étude propose une mesure de la littératie financière adaptée pour tenir compte de la longueur de l’enquête.

Sujets : Méthodes économétriques, statistiques et computationnelles; Argent comptant (billets de banque); Actifs numériques et technologies financières; Infrastructures de paiement et de marchés financiers; Paiements de détail
Codes JEL : C81, C83, D12, G53

1. Introduction

Financial literacy is commonly measured with the "*Big Three*" questions developed by Lusardi and Mitchell (2011). These questions assess respondents' understanding of the concepts of compound interest rates, inflation, and risk diversification. These elements are widely used to construct indicators of financial literacy and to study financial decision making; see Lusardi and Mitchell (2008, 2011); van Rooij et al. (2011). Lusardi and Mitchell (2014) highlight that the level of financial literacy is quite low in a set of countries (Australia, France, Germany, Italy, Japan, Netherlands, New Zealand, Romania, Russia, Switzerland, and the USA). Financial literacy questions are usually part of a broader survey that is lengthy. Valdes et al. (2024) note that the American National Financial Capability Study (NFCS) run over a twelve-year period (2009-2012) is a succinct survey and every effort was made to reduce respondent fatigue and measurement error, may contain measurement errors.

In this paper, we investigate the role of survey fatigue as a potential mechanism for low financial literacy scores. Herzog and Bachman (1981) find that an increase in survey length and induce survey fatigue. This survey fatigue can lead to lower quality responses, particularly for questions placed later in the instrument. Therefore, we use the A/B experiment design, where survey respondents are randomly assigned to receive the financial literacy questions either at the beginning (hereafter known as the A sample) or near the end of the survey (hereafter known as the B sample). This experimental A/B test was embedded into the Cash Alternative Survey (CAS) conducted in November 2020 by the Bank of Canada. This design allows for a direct assessment of how placement affects response patterns.

We find that placing financial literacy questions at the end of the survey increases the responses of "Don't know" (DK) and reduces the correct answers by 5% to 15%, with little impact on the rate of incorrect answers. Respondents were more likely to provide satisficing such as DK or expend less effort on cognitively demanding items when fatigued; see Krosnick et al. (2002). Jeong et al. (2023) find a similar result when they randomize the order of questions in lengthy in-person surveys or phone surveys. They find that an increase of one to two hours in survey length will lead to an increase in the probability of a respondent skipping a question by 10% to 64%. Valdes et al. (2024) find that in the NFCS, financial literacy as measured by the Big 5 has steadily declined over a 12-year period (2009-2021). This decline is associated with a rise in the percentage of respondents answering DK for the questions. Followup research by Burke et al. (2026) analyze the longitudinal data and find similar result to us that respondents were likely to answer DK than correct responses. Further, they conduct a randomization of the survey mode (computer, smartphone, and tablet) as a potential reason for the increase in DK responses. They attribute 23% of the decline in financial literacy is due to smartphone as a survey mode.

We use Oaxaca–Blinder decompositions, as discussed in Blinder (1973); Oaxaca (1973), highlight that the effect is especially pronounced among respondents who are women, younger adults, or those with lower educational attainment. The probability of a DK response rises for high-school-educated respondents when financial literacy questions are placed later in the survey, while university-educated respondents are less affected. Age also matters; the youngest group starts out with the highest rate of DK responses, and this group's DK rate increases the most when survey fatigue sets in. Lusardi and Mitchell (2023) review the cross-country evidence and find differences in financial literacy based on demographics. For example, women tend to score lower in financial literacy, partly because women are more likely to answer questions with DK. Experimental evidence by Bucher-Koenen et al. (2025) finds that once the DK is removed, the gap between men and women is lower. They attribute a substantial share of the gender gap can be traced to differences in confidence. Hospido et al. (2024) use a multi-arm randomized control trials to disentangle the removal of the DK option, monetary incentives, and informational nudge. They find that the information nudge can significantly reduce the gender gap in choosing DK and in financial literacy.

In summary, a key implication of our research is that rates of incorrect responses remain stable across experimental groups, while correct and DK responses are sensitive to survey structure. This suggests that incorrect answers may serve as a more robust benchmark for financial literacy comparisons across surveys with differing designs. These results are consistent with Delavande et al. (2008), who find that survey fatigue factors in the amount of time that a respondent will devote to answering complex financial questions. If they have low knowledge of financial concepts, then they are more likely to answer DK.

The remainder of the paper is structured as follows. Section 2 describes the survey data used in the study, Section 3 discusses the methodology and the results, Section 5 reports robustness checks, and Section 6 concludes. Additional analyses and discussion are provided in the Appendix.

2. November 2020 Cash Alternative Survey

The primary data used in this analysis comes from the CAS; see Chen et al. (2021). This consumer survey was conducted in November 2020 by the Bank of Canada to monitor the use of payment methods by Canadians during the second wave of the COVID-19 pandemic. The motivation to undertake the survey randomization was to test the efficacy of the placement of the financial literacy questions. At the beginning of the pandemic, the April 2020 CAS found that financial literacy had fallen slightly; see Chen et al. (2020), vis-á-vis Huynh et al. (2020) and Henry et al. (2018). In addition, the November 2020 CAS (about 50 questions) was lengthened relative to the April 2020 CAS (about 25 questions). The format of the November 2020 CAS was similar to earlier Methods-of-Payment (MOP) surveys conducted by the Bank of Canada in 2009, 2013, and 2017. The November 2020 CAS serves as the benchmark design for the later MOP surveys, conducted from 2021 to 2024. In 2021-2023, the financial literacy questions appeared earlier on in the survey, see Henry et al. (2022, 2024a,b). In 2024, the financial literacy questions were moved to a later on in the survey given this research, see Felt et al. (2025).

2.1. Survey Design

The November 2020 CAS respondents were required to complete two tasks: a roughly 20-minute survey questionnaire (SQ) and a three-day diary survey instrument (DSI). In the SQ, they responded to questions about cash holdings, situations in which cash was not used, the use of credit cards and other forms of payment, and perceptions about types of payment in general, among others. Respondents were also asked to answer the "Big Three" financial literacy (FL) questions (Lusardi and Mitchell, 2011; **?**), which assess a respondent's comprehension of compound interest, inflation, and risk diversification.

The November 2020 CAS followed a random A/B test design; see Kohavi et al. (2009), where survey panelists were assigned to one of two groups based on the placement of their FL block. Respondents who received a version of the questionnaire with the FL questions at the beginning were denoted as being in Group A (the "untreated" group) while respondents receiving the FL questions close to the end were considered to be in Group B (the "treated" group). Appendix A.1 presents a visual of the survey flow (see **Figure A-1**). Other than the placement of the FL questions, the questionnaire structures were identical for all respondents. After the SQ questions, respondents were asked to fill in the DSI to log their payment and withdrawal activities over the course of three days. In the CAS, 3,893 participants completed the SQ, 2,084 of whom also completed the diary.

2.2. Survey Methodology

Respondents were recruited with non-probability quota sampling. To align with certain Canadian demographics, survey weights were subsequently constructed. More information on the construction of weights can be found in Chen et al. (2021). We use these weights where appropriate in our analysis. The Bank of Canada collaborated with Statistics Canada to address potential concerns of the non-probability quota sampling. This entailed benchmarking the November 2020 CAS to the Canadian Perspectives Survey Series 5: Technology Use and Cyber Security During the Pandemic Public Use Microdata File (a probability survey conducted in September 2020, hereafter known as CPSS5); see Statistics Canada (2020a). That is, the Bank of Canada included questions from the CPSS5 about consumers' precautions during COVID-19 and their usage of technology during this period of time.

The CPSS5 was chosen for the following reasons: first, the CPSS sample (number of observations was 3,961) comes from rotation groups of the Labour Force Survey, a reliable social probability survey. Second, the CPSS5 data was collected from September 14, 2020, to September 20, 2020, similar to the field operation period for the November 2020 CAS. In addition, both the Statistics Canada CPSS5 and the Bank of Canada November 2020 CAS were conducted in the online mode.

Further, Chen and Tsang (2026) develop a statistical inferences method for a non-probability two-phase survey sample when relevant auxiliary information is available from a probability survey sample. To reduce selection bias and gain efficiency, both selection probabilities of Phase 1 and Phase 2 are estimated, and two-phase calibration is implemented. They find negligible differences between the non-probability quota sampling and the probability survey sample from Statistics Canada. For robustness, we account for differences in the November 2020 CAS and the CPSS5 in the results. Similar to Chen and Tsang (2026), we do not find any economic or statistical differences due to using a non-probability sampling.

3. Econometric Analysis of the Randomization

In this section, we undertake an econometric analysis of randomization. We first start with descriptive statistics followed by conditional analysis via regression analysis. Finally, we conclude with decomposition of the differences using Oaxaca–Blinder decompositions; see Oaxaca (1973); Blinder (1973).

3.1. Descriptive Statistics

We first provide an overview of the November 2020 CAS, as well as some descriptive statistics for the two treatment groups (A and B). We can write the random assignment of a respondent to Group A versus Group B as:

$$T_i = \begin{cases} 0 & \text{if participant } i \text{ receives block at beginning of survey (A)} \\ 1 & \text{if participant } i \text{ receives block at end of survey (B),} \end{cases} \tag{1}$$

where i is a participant and T_i is the treatment indicator. The November 2020 CAS has the same number of observations in groups A and B ($n_A = n_B = 1,946$).

First, as a simple test, we construct a balance table to compare a number of observable demographics across groups A and B. **Table 1** reports the weighted frequencies of respondents with the following characteristics: gender (male, female), age (18-34, 35-54, 55+), highest educational attainment (high school, college, university), region (British Columbia, Prairies, Ontario, Quebec, Atlantic), income (low or less than Can$45K, medium or between Can$45K and Can$85K, and high or above Can$85K), marital status (unmarried, married), and children living at home (has children at home, does not have children at home). We note that demographic shares are balanced, with no more than a 3-percentage-point gap between the shares of groups A and B.

The financial literacy questions used are the "Big Three," presented in a standard order that remains the same for both Group A and Group B (see Appendix section A.1, **Table A-1**). Respondents can answer each question one of three ways: by selecting the correct response, an incorrect response, or answering "Don't know." **Tables 2, 3,** and **4** show the weighted shares of each response type, by treatment group, question, and select demographic groups. Already, it appears that responding to financial literacy questions at the end of the questionnaire is associated with a lower proportion of correct answers (**Table 2**) and a higher proportion of DK responses (**Table 4**), and this seems to apply to all demographic groups. **Table 3** shows that the trends in incorrect responses are not quite as salient across demographic groups. For example, for the inflation question, some demographics (such as men) have a lower incorrect response rate in Group B, while others (such as women) have a lower rate in Group A. For the interest rate and risk questions, Group B tends to have a higher incorrect response share, though this difference is not always large.

The results suggest that, in general, Group A outperforms Group B, and this appears to be attributed to Group B respondents answering correctly less frequently and DK more frequently. Furthermore, we note that groups A and B appear to be balanced in observables, and both groups contain an identical number of respondents. In the following subsections, we conduct a conditional analysis to understand potential demographic factors with the results.

3.2. Benchmark Comparison Based on Random Assignment

To quantify the difference in correct, incorrect, and DK responses for individual questions, we seek to compute the difference in responses between the two groups:

$$\Delta Y_{ij} = P(Y_{ij} = 1 | T_i = 1) - P(Y_{ij} = 1 | T_i = 0), \tag{2}$$

where $Y_{ij} \in \{Correct_{ij}, Incorrect_{ij},\ DK_{ij}\}$, which are indicators taking the value 1 if the respondent gives a correct, incorrect, or DK response (respectively) to question j. Our goal is to estimate the difference in response probabilities for a characteristic respondent i, the treatment effect of receiving the FL block later in the survey. As respondents can only belong to one of the two treatment groups, we cannot compute this difference directly.

However, we can estimate the difference in a number of ways. Our first method is by using linear regression:

$$Y_{ij} = \beta + \gamma Ti + u_{ij}, \tag{3}$$

where $Y_{ij} \in \{Correct_{ij}, Incorrect_{ij}, DK_{ij}\}$ (the observed indicators) and γ is a measure of the share difference (e.g., an average $\Delta Correct_j$ across individuals). We can then add demographic controls to this regression, to account for potential confounding effects.

One key assumption is the zero conditional mean assumption: the expectation of the error term in this regression, conditional on treatment, $E[u_{ij}|T_i]$, is zero for both treatment groups. Due to the random assignment of treatment, we have no a priori reason to suspect that this would not be the case. We perform several robustness checks in Section 5.

We conduct a nonparametric method to test if the response distribution differs among the groups by using a χ^2 test of independence. The aim is to check whether the distribution of overall responses differs significantly between groups A and B. For each question, we set up a 2×3 contingency table, with treatment groups as the rows and the response types as columns. Using the cell counts, we test the null hypothesis that response categories of groups A and B follow an identical distribution. This approach and its results are further documented in Appendix section A.2.

3.2.1. Regression Analysis

We begin by examining the difference in response type shares. As presented in **Table 5**, Group B consistently achieved lower correct response rates in all three FL questions. The gap between groups A and B persists even when accounting for the standard set of demographic characteristics: gender, age, education, region, income, marital status, and having kids in the home. **Figure 1** visually represents the gap in correct, incorrect, and DK response shares. The largest difference we observe is in the risk diversification question, where the treated group has a 15-percentage-point decrease in the correct answer rate, and a corresponding 14-percentage-point increase in the DK response rate (using all controls). The gap in incorrect responses is significant only for the interest rate question, where the treated group has a 3-percentage-point increase in incorrect responses, relative to Group A.

In the inflation and risk questions, it appears that Group B's lower correct response rates are associated with an increased prevalence of DK, while incorrect responses remain relatively stable. We suspect that the mechanism underlying this phenomenon is survey fatigue.

3.2.2. Conceptual Framework

We provide a conceptual framework based on Delavande et al. (2008) to assist in understanding these results. Consider that there is a cost to attempting an FL question j. This cost imposes a cognitive cost, c_{ij}, on a respondent i. This cost is unobserved, could vary for respondents based on observable and unobservable characteristics, and is higher at the end of a survey when a respondent is tired or bored. There is no cost for not attempting question j, and so respondents may choose to opt out by either guessing or selecting DK. The benefit to attempting question j is also unobserved, but it is assumed this exists and is higher for correctly answering a question. The benefit could be receiving survey compensation, for instance, especially if respondents believe opting out of these questions may lead their responses to be discarded and compensation to be withheld.

Each individual respondent will have a cost threshold, where a c_{ij} above the threshold will induce them to expend no effort on the question, and a c_{ij} below will induce them to expend effort in an attempt to correctly answer the question. This threshold could reflect a number of characteristics, including confidence, the prior belief they can correctly answer the question, or how busy the respondent is. Respondents with a low threshold will choose not to engage with question j regardless of whether it is placed at the beginning or end of the survey. However, some respondents may have a cost threshold inducing them to expend effort at the beginning of the survey, but not at the end. These respondents could choose to either randomly guess an answer or to select DK.

If a high proportion of these respondents have low FL (would fail to correctly respond to question j if attempted), we might observe a similar share of correct responses in groups A and B, a similar or lower share of incorrect responses in Group B, and a higher share of DK responses. The exact sign of incorrect responses would depend on whether respondents choose to randomize or to select DK when exerting no effort—findings from Bertola and Lo

Prete (2025) suggest that most individuals with low FL would choose to randomize, especially for easier questions. However, if a high proportion of these respondents have high FL (can correctly respond to question j when attempted), then we would observe a lower share of correct responses in Group B and a higher share of incorrect and DK responses.

The results suggest that some respondents have low FL, while others have high FL and choose to exert no effort. Most likely, a slightly larger share of respondents have true high FL, as in Bucher-Koenen et al. (2025). This would explain why we observe an overall increase in DK responses, and an almost-corresponding decrease in correct responses. The interest rate question, which is considered the easiest question, is one where the second case may hold more than the first, hence why both incorrect responses (+3 percentage points) and DK responses (+4 percentage points) increase in Group B. The next section investigates the role of observables and unobservables.

3.3. Oaxaca–Blinder Decomposition

In addition to measuring the gap between the FL answers in the two surveys, we are also interested in understanding how observable and unobservable characteristics help us explain this gap. To do this, we consider the size of the treatment gap using what is known as a Oaxaca–Blinder (OB) (Oaxaca, 1973; Blinder, 1973) decomposition, which quantifies how much of the gap can be explained and how much unexplained by different factors.

We primarily use a pooled model (Oaxaca and Ransom, 1994), which includes the use of a non-discriminatory coefficient β. We use the Stata *oaxaca* package developed by Jann (2008), with the pooled option to conduct the decompositions. The mean difference in FL outcomes by treatment ($\Delta Y = E(Y_B) - E(Y_A)$) can be decomposed as follows:

$$\Delta Y = [E(X_B) - E(X_A)]'\beta + E(X_A)'(\beta - \beta_A) + E(X_B)'(\beta_B - \beta), \tag{4}$$

where A is the subsample of respondents in Group A, B is the subsample in Group B, and $E(Y_g) = \beta_g E(X_g)$ by subgroup $g \in \{A, B\}$.

The goal of this methodology is to uncover any portions of the observed FL gap that are attributable to differences in observables between Groups A and B. It also helps us to identify which subgroups of respondents may be driving the gap, which we can then further investigate.

First, the OB decomposition controls for endowments (explained observed characteristics), consisting of gender, age, education, region, income, marital status, and children living at home. The endowments component shows us to what extent differences in FL scores are simply due to the compositional differences between the groups. For instance, if the average educational level of Group A is higher than that of Group B, the endowments component will measure the extent to which the education gap between the two groups is associated with differences in FL scores:

$$[E(Educ_B) - E(Educ_A)]'\beta. \tag{5}$$

Second, OB accounts for returns to unobserved characteristics (coefficient). This term represents the role of unobservable characteristics in the differences in FL. This is especially important in our case of A/B testing. A large coefficient component indicates that the influence of covariates on FL may be different across survey design groups. For instance, suppose groups A and B have identical education levels ($E(Educ_A) = E(Educ_B) \equiv E(Educ)$); the coefficient part measures whether higher education is"rewarded" more (in terms of FL outcomes) when Group A has higher education relative to Group B:

$$E(Educ)'(\beta_B - \beta_A). \tag{6}$$

In our experimental design, this may signal some form of survey fatigue affecting one group more than the other. Therefore, the OB analysis does tell us about subgroups that could be susceptible to survey fatigue. The coefficient component also comprises the constant, i.e., the intercept on group variance in FL.

Unobserved factors might play a role in our analysis in several ways. First, and perhaps more problematically, are non-fatigue-related survey design effects, such as bias of survey attrition by group. If respondents in one group leave the survey earlier or more often than respondents in the other, then our estimates of the fatigue effects may be biased: respondents in one group would come from a population more or less willing to fill the survey.

Alternatively, there could still be differences in attrition due to placement of FL questions if, for example, respondents drop out shortly after being exposed to the FL questions because they are discouraged by those FL questions. Other unobservables that enter into our analysis are potential psychological factors, including survey fatigue, boredom, test-taking specific behavior (i.e., submitting answers with the highest probability of being correct rather than most likely), and social desirability bias (responding in a manner that is more socially acceptable). Differences in FL can also be due to differences in cognitive ability, personality characteristics, financial attitudes, access to financial information, and past experiences with money. We would add that the presence of a substantial unexplained component can be viewed, indicating either a survey fatigue effect or potential biases in our specification.

The OB decomposition builds on the above findings by decomposing the observed group gaps into two components: differences in endowments (observable characteristics such as age and education) and differences in coefficients (how these characteristics differentially contribute to outcomes). The OB results confirm the main findings: in particular, that Group B underperforms relative to Group A.

The decomposition results add another layer of analysis by showing that the observed differences between the two groups are almost entirely due to differences in coefficients rather than [in?] endowments. This means that the difference is not about who is in each group, but how the survey context (i.e., question placement) affects the way various subgroups answer. For example, all subgroups, even educated respondents (who tend to show higher FL based on responses to the "Big Three" survey questions), are less likely to answer correctly and more likely to answer DK when FL questions are placed at the end. This highlights a context-driven effect rather than a compositional one. OB decomposition allows us to identify some of the subgroups that may be more or less impacted by survey fatigue, which we further explore in Section 4.

In the interest rate question (**Table 6**), treatment still results in an 8-percentage-point decrease in correct responses, a 3-percentage-point increase in incorrect responses, and a 5-percentage-point decrease in DK answers. The endowments component does not significantly or numerically contribute to this gap, with most of the difference coming from the coefficients component. Significant coefficient covariates include education, region, and gender. Differences in the constant also contribute somewhat to the gap.

Table 7 shows results of the inflation question. As with the interest rate question, results are similar to the primary analysis. Treatment results in a 6-percentage-point decline in correct responses, an insignificant 2-percentage-point decline in incorrect responses, and a 7-percentage-point increase in DK responses (relative to Group A). As before, the endowments component plays an insignificant role. Significant contributions to the gap come from within-group coefficient differences in region and educational attainment.

Finally, the largest gap for the correct and DK answers is found for the risk question (**Table 8**). Here we observe a 16-percentage-point drop in correct answers from Group A to Group B and a 15-percentage-point increase in the DK answers. This result also emphasizes the difficulty of the risk question for the respondents. The significant coefficient contributions come from age, education, income, region, and the constant.

4. Heterogeneity Analysis

Lusardi and Mitchell (2014) survey the literature and find that several demographic groups (gender, age, and education) are identified as lagging in FL. De Bassa Scheresberg (2013) examines FL among young adults, finding that young women, minorities, and less-educated individuals demonstrate lower FL scores based on responses to survey questions. While the existence of demographic differences in FL is well documented, we investigate whether survey fatigue could contribute to some of these differences.

From the decomposition, we focus on three demographic groups that appear as significant coefficient covariates: gender, age, and education. In addition, income and region are also found to occasionally contribute to financial literacy gaps. We do not investigate these separately, as income likely correlates with the other subgroups (i.e., age, education, and gender) and we are less focused on estimating regional differences in financial literacy. In the appendix, we conduct further analysis conditioning on an indicator for uncertainty at other points in the survey; see Appendix section A.6.

To test the differential treatment effect for each subgroup, we first estimate an interaction specification model:

$$Y_{ij} = \beta + \gamma T_i + \alpha S_i + \delta(T_i \times S_i) + \eta X_{ij} + u_{ij}, \tag{7}$$

where S_i is an indicator for the subgroup of interest (e.g., gender), δ is an estimate of the heterogeneous treatment effect, and X_{ij} are standard demographic controls (age, gender, education, income, having children [in the home?], marital status, province). Y_{ij} is an outcome variable for individual i and question $j \in \{Interest, Inflation, Risk\}$. It is either the total count of a respondent's correct, incorrect, or DK answers, or it is an indicator for whether they answered any of the three financial literacy questions correctly, incorrectly, or as DK.

We also consider two alternative specifications, using logistic regression for the indicator outcomes and ordered logistic regression for the limited count.[1] Our results examine how Group B treatment interacts with certain demographic variables, allowing us to observe whether certain respondents may be more impacted by survey fatigue than others.

4.1. Gender

The results of the conditional treatment effect on gender are presented in **Table 9**. The interaction effect between gender and treatment is not significant for the number of correct and incorrect responses and only marginally significant for the number of DK responses. In this case, we identify that women in Group B have an additional 12% of DK responses relative to men who are treated. It is important to note that this is largely driven by the interest question, which is the only question that identifies gender as significant in the OB decomposition.

The intensive margin is also driving this effect: conditioning on respondents answering DK at least once, we find that *only* the female × treatment interaction is significant, and the level treatment effect disappears entirely. Regressing an indicator for answering DK to at least one FL question on treatment, gender, and the interaction, we find no significant effects. Therefore, it appears that while women are not disproportionately affected by survey fatigue at the extensive margin of responding DK, placing FL questions later in the survey may induce them to provide *more* DK answers relative to men if they are already providing at least one DK response.

One possible explanation for this is that women who already lack confidence in their FL responses will be disproportionately affected by fatigue. Bucher-Koenen et al. (2025) find that 30% of the gender gap in "Big Three" responses is attributable to a lack of confidence. The roles of experience, confidence, and gender differences are explored in Balutel et al. (2023). They find no significant gender differences in FL among Bitcoin owners. However, for non-owners of Bitcoin, there are gender differences in both crypto and financial literacy. Our results suggest that survey fatigue could potentially compound the effect of lack of confidence. However, more investigation is required to disentangle confidence from survey fatigue.

[1] Ordered logistic regression is typically used when an outcome consists of ordered categories, but in this case, we consider it for the number of correct, incorrect, or DK responses.

4.2. Age

Table 10 shows the interaction regressions using age categories. Throughout our analysis, we code age into three distinct categories to account for potential nonlinearities. Our omitted category is the youngest age group, consisting of 18- to 34-year-olds. We interact the other two age categories with treatment. Altogether, respondents in Group B answer 0.20 fewer questions correctly, and respond DK to 0.24 more questions. At baseline, younger respondents perform worse than older respondents. The 35- to 54-year-olds have 0.34 more correct responses ($p < 0.01$), 0.29 fewer incorrect responses ($p < 0.01$), and a similar number of DK answers (0.06 fewer, $p > 0.1$) relative to those in the 18-34 age group. Respondents aged 55+ have an even larger number of correct responses (0.67 more, $p < 0.01$), fewer incorrect ones (0.41 fewer, $p < 0.01$), and fewer DK answers (-0.27, $p < 0.01$).

However, the interaction effects are, for the most part, small or insignificant, meaning that we cannot reject the null hypothesis that all three age categories are impacted equally by survey fatigue. The 35- to 54-year-olds in the treated group answer an additional 0.14 questions incorrectly, but this is only marginally significant ($p < 0.1$). The 35- to 54-year-olds also appear to answer 0.17 fewer questions correctly when treated, but this estimate is insignificant.

Ultimately, there is no strong evidence that survey fatigue impacts different age groups differently, though it appears that the middle-aged group performs worse when faced with fatigue—answering more questions incorrectly. One explanation for this could be that these respondents are overwhelmed with work, children, or busy [lives/schedules], and so they expend less cognitive effort on the FL questions while still feeling confident in the answer.

4.3. Education

Table 11 shows the interaction regressions using education categories. Regardless of treatment status, we note that college- and university-educated respondents perform much better than respondents who have completed only high school, our baseline group. College-educated respondents have on average 0.23 more correct responses ($p < 0.01$), a similar number of incorrect responses (0.08 fewer; $p > 0.1$), and 0.15 fewer DK responses ($p < 0.01$) than respondents who have completed only high school. University-educated respondents have 0.46 more correct responses ($p < 0.01$), 0.24 fewer incorrect responses ($p < 0.01$), and 0.22 fewer DK responses ($p < 0.01$).

Estimating the interaction effects suggests university graduates in particular are less prone to growing fatigued. The overall treatment effect is 0.33 fewer correct responses ($p < 0.01$), 0.01 more incorrect responses ($p > 0.1$), and 0.32 more DK responses ($p < 0.01$). The interaction of treatment and having a college education leads to small and insignificant deviations. However, the interaction between university and treatment reveals that the treatment effect for these individuals is 0.16 fewer correct options (marginally significant with $p < 0.1$), and 0.12 more DK responses ($p < 0.01$). In other words, the estimated treatment effect for individuals with the highest education is around half of what it is for individuals with only a high school diploma.

This could be explained by university graduates having more familiarity with financial concepts due to their education. De Bassa Scheresberg (2013) finds that correct responses to the "Big Three" questions is increasing in education, which supports this idea. With more familiarity, university graduates may expend less cognitive energy on the FL questions, making those respondents less prone to fatigue. For similar reasons, they could also be more comfortable with the multiple-choice style of question.

5. Robustness

While our results support the hypothesis that survey fatigue induces more respondents to answer DK on the FL questions, unobservable factors could still play a role. For instance, there could be higher survey attrition in Group A, if respondents who are less confident in their financial knowledge (or who are already fatigued) become overwhelmed when encountering the FL questions. Surviving Group A respondents could, in this case, outperform Group B in the FL questions by default. Furthermore, the quota sampling of the CAS could impact the A/B randomization. To test the validity of our results, we perform a number of checks.

5.1. Analyzing Consumer Cash Holdings

As an additional test, we investigate the impact of FL placement on a cash holdings question placed in the middle of the survey for both groups. Respondents are asked to report their cash holdings, both in terms of the amount of cash on hand or in their wallet, and in terms of any other cash holdings they may keep. Given the hypothesis that survey fatigue drives our main results, we would not expect to see different outcomes between Group A and Group B for questions such as the cash holdings question. If fatigue does affect respondents for this question, we would expect it to affect both groups similarly.

Figure 2 shows the distributional differences in cash holdings (on hand and other cash holdings) for the two groups. Figure 2 also shows that, across the two groups, cash is overlapping in distribution for both cash on hand and other cash. Table 12 shows the results of the OB decomposition, using the cash on hand and other (external) cash holdings, in dollars of CAN$, as dependent variables.

In each case, we do not observe a large or significant gap between groups A and B. For the unaltered cash holdings variables, we also do not notice any significant endowment or coefficient covariates. However, when we instead use the logarithm of cash holdings, some insights are revealed. The impact of income on cash holdings may be nonlinear, with education playing a significant role in how these variables interact. The unexplained variance in cash holdings appears to be greater among higher-income or college-educated individuals. The analyses on cash holdings in general suggest that placement of FL questions does not directly impact how other personal-finance questions are answered, though it may port some interaction effects that are present and can be captured in the unobservables.

Ultimately, the lack of significant differences in cash holdings suggests that placement of the FL questions does not have a direct or substantial impact on other survey sections. In other words, there is no carryover effect strong enough to influence responses to subsequent unrelated questions about personal finance.

5.2. Benchmarking to a Probability Survey

There may be concerns that our study uses a non-probability quota sampling and that the representativeness of the November 2020 CAS survey sample. Establishing external validity is essential for generalizing our findings beyond the sampled respondents even considering population weights; see (Chen et al., 2021). To address this, we leverage Statistics Canada probability survey data from 2020, specifically, the Canadian Perspectives Survey Series 4 and 5 (Statistics Canada, 2020b,a) to benchmark the November 2020 CAS sample against national demographic and behavioral trends.

Our approach focuses on 14 survey questions that are identical in both the November 2020 CAS and Statistics Canada CPSS5 datasets. These questions cover COVID-19-related behaviors (such as mask-wearing and pandemic preparation) and internet security practices (such as password strength and online shopping habits), all coded as binary (Yes/No) variables, denoted as S. To assess representativeness, we first calculate the weighted means for each question by age and gender group (18 to 34, 35 to 54, 55+; male, female) in the Statistics Canada data, denoted as $\bar{S}^{STC}_{A,G}$. For each CAS respondent, we then compute a "detrended" score, measuring the deviation from their probability-sampled national average. Using the CAS data, we then detrend individual responses, given their age and gender group, to get a measure of deviation from probability-sampled expected values:

$$\tilde{S}_i = S^{CAS}_i - \bar{S}^{STC}_{A,G}. \tag{8}$$

This metric captures the differences between each November 2020 CAS respondent's answer relative to their demographic group in the CPSS5 representative sample. Next, we regress these detrended scores ($\tilde{S}_i$) on the full set of CAS covariates: age, gender, region, education, parental status, and marital status, but exclude treatment assignment:

$$\tilde{S}_i = \omega X_i + \epsilon_i. \tag{9}$$

The estimated error term, $\hat{\epsilon}_i$, represents the portion of each respondent's deviation from the national mean that cannot be explained by standard observable characteristics. These error terms serve as proxies for any unobservable differences between the CAS and Statistics Canada respondents.

To further test whether these unobservable differences might bias our treatment effects, we incorporate the predicted error terms ($\hat{\epsilon}_i$) into the OB decompositions. This allows us to determine whether the observed differences in treatment groups in financial literacy outcomes are driven by unmeasured sample divergence. We consider the total sum of errors. We can also consider each predicted error term as a separate variable in the decomposition.

Our findings (see **Tables 13, 14**, and **15**) indicate that including these predicted error measures only slightly increases the explained share in OB decomposition. The endowment portion of the difference becomes marginally significant for most questions and responses, too, though it is still small in magnitude. This suggests that the unexplained part of the individual-level CPSS deviation could possibly differ slightly by treatment group, though not largely.

A possible explanation for the slight difference between groups could be the placement of the CPSS questions themselves. In the CAS, they are the second-last set of questions in the survey, placed soon after the FL questions appear for Group B. Group B may be tired, having just seen the relatively difficult FL questions, perhaps inducing them to expend less effort on the CPSS questions compared with Group A.

However, including the sums of CPSS residuals does not change our results. Deviations from the Statistics Canada means do not meaningfully account for the treatment effect observed in our main analysis. These results are similar to Chen and Tsang (2026) find negligible differences for the cash volume shares.

6. Conclusion

This study provides robust evidence that survey design, specifically the placement of the financial literacy questions, significantly influences the accuracy of the respondents and their likelihood of selecting DK versus correct answers. Our findings reveal a consistent pattern: when the "Big Three" financial literacy questions are placed at the end of a survey, there is a notable decrease in correct answers and a corresponding increase in DK responses. This points to survey fatigue as a key factor influencing DK responses.

The Oaxaca–Blinder decomposition confirms that these differences are primarily driven by question placement rather than underlying differences in respondent characteristics, and we demonstrate that this effect is particularly pronounced for women, certain age groups, and less-educated individuals.

An especially important finding is the role of incorrect answers in the FL questions. In contrast to the sensitivity of correct and DK responses to survey context, we find that incorrect answers remain remarkably stable across treatment groups. This suggests that indices of financial literacy should be grounded primarily in incorrect responses rather than in correct answers, as relying on correct responses may introduce bias and distort true knowledge levels due to their variability. Emphasizing incorrect answers as the basis for literacy indices ensures a more consistent and valid assessment of financial knowledge, ultimately improving the accuracy of identifying knowledge gaps and informing effective financial education and policy interventions.

Furthermore, understanding how DK responses function in financial literacy surveys can lead to more accurate assessments of population knowledge levels, inform targeted educational interventions, and ultimately contribute to improving financial decision-making capabilities in an increasingly complex economic environment. Therefore, we advocate for a greater emphasis on incorrect responses as a more stable and valid measure of financial literacy.

References

Ai, C., E. C. Norton, and H. Wang (2004): "Computing interaction effects and standard errors in logit and probit models," *The Stata Journal*, 4, 154–167.

Balutel, D., W. Engert, C. S. Henry, K. P. Huynh, D. Rusu, and M. C. Voia (2023): "Crypto and Financial Literacy of Cryptoasset Owners Versus Non-Owners: The Role of Gender Differences," *Journal of Financial Literacy and Welbeing*, 1, 514–540.

Bertola, G. and A. Lo Prete (2025): "Who prefers guessing to admitting They Don't Know? Measurement error in financial literacy surveys," *Journal of Economic Behavior and Organization*, 233.

Blinder, A. S. (1973): "Wage discrimination: Reduced form and structural estimates," *Journal of Human Resources*, 8, 436–455.

Bucher-Koenen, T., R. Alessie, A. Lusardi, and M. van Rooij (2025): "Fearless Woman: Financial Literacy, Confidence, and Stock Market Participation," *Management Science*, 71, 7414–7430.

Burke, J., C. Urban, and O. Valdes (2026): "Is Financial Knowledge Really Declining? Randomized Evidence on the Effects of Smartphone Responses," Working Paper 2026-004, CESR-Schaeffer Working Paper Series, funded by the FINRA Investor Education Foundation.

Chen, H., W. Engert, M.-H. Felt, K. P. Huynh, G. Nicholls, D. O'Habib, and J. Zhu (2021): "Cash and COVID-19: The impact of the second wave in Canada," Staff Discussion Papers 2021-12, Bank of Canada.

Chen, H., W. Engert, K. P. Huynh, G. Nicholls, M. Nicholson, and J. Zhu (2020): "Cash and COVID-19: The impact of the pandemic on demand for and use of cash," Staff Discussion Papers 2020-6, Bank of Canada.

Chen, H. and J. Tsang (2026): "Correcting Selection Bias in Non-Probability Two-Phase Payment Survey," *Journal of Survey Statistics and Methodology*, smaf043.

De Bassa Scheresberg, C. (2013): "Financial Literacy and Financial Behavior among Young Adults: Evidence and Implications," *Numeracy*, 6.

Delavande, A., S. Rohwedder, and R. Willis (2008): "Preparation for Retirement, Financial Literacy and Cognitive Resources," Working Papers wp190, University of Michigan, Michigan Retirement Research Center.

FELT, M.-H., A. CHERNESKY, AND A. WELTE (2025): "2024 Methods-of-Payment Survey Report: Cash in an Era of Alternatives," Staff Discussion Papers 2025-12, Bank of Canada.

HENRY, C. S., K. P. HUYNH, AND A. WELTE (2018): "2017 Methods-of-Payment Survey Report," Staff Discussion Papers 2018-17, Bank of Canada.

HENRY, C. S., D. RUSU, AND M. SHIMODA (2024a): "2022 Methods-of-Payment Survey Report: Cash Use Over 13 Years," Staff Discussion Papers 2024-01, Bank of Canada.

HENRY, C. S., M. SHIMODA, AND D. RUSU (2024b): "2023 Methods-of-Payment Survey Report: The Resilience of Cash," Staff Discussion Papers 2024-08, Bank of Canada.

HENRY, C. S., M. SHIMODA, AND J. ZHU (2022): "2021 Methods-of-Payment Survey Report," Staff Discussion Papers 2022-23, Bank of Canada.

HERZOG, A. R. AND J. G. BACHMAN (1981): "Effects of Questionnaire Length on Response Quality," *The Public Opinion Quarterly*, 45, 549–559.

HOSPIDO, L., N. IRIBERRI, AND M. MACHELETT (2024): "Gender gaps in financial literacy: a multi-arm RCT to break the response bias in surveys," Working Papers 2401, Banco de Espana.

HUYNH, K. P., G. NICHOLLS, AND M. NICHOLSON (2020): "2019 Cash Alternative Survey Results," Staff Discussion Papers 2020-8, Bank of Canada.

HUYNH, K. P. AND M. C. VOIA (2024): "Don't Know! Don't Care? We Should! Gender Differences on "Don't Know" Responses in Digital and Financial Literacy Questions," *mimeo.*

IMAI, K., G. KING, AND E. A. STUART (2008): "Misunderstandings Between Experimentalists and Observationalists about Causal Inference," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171, 481–502.

JANN, B. (2008): "The Blinder-Oaxaca decomposition for linear regression models," *The Stata Journal*, 8, 453–479.

JEONG, D., S. AGGARWAL, J. ROBINSON, N. KUMAR, A. SPEAROT, AND D. S. PARK (2023): "Exhaustive or exhausting? Evidence on respondent fatigue in long surveys," *Journal of Development Economics*, 161, None.

KOHAVI, R., R. LONGBOTHAM, D. SOMMERFIELD, AND R. M. HENNE (2009): "Controlled experiments on the web: survey and practical guide," *Data Mining and Knowledge Discovery*, 18, 140–181.

KROSNICK, J. A., A. L. HOLBROOK, M. K. BERENT, R. T. CARSON, W. M. HANEMANN, R. J. KOPP, R. C. MITCHELL, S. PRESSER, P. A. RUUD, V. K. SMITH, ET AL. (2002): "The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice?" *Public Opinion Quarterly*, 66, 371–403.

Lusardi, A. and O. S. Mitchell (2008): "Planning and financial literacy: how do women fare?" *American Economic Review*, 98, 413–417.

——— (2011): "Financial literacy and planning: implications for retirement wellbeing," *Journal of Pension Economics and Finance*, 10, 497–508.

——— (2014): "The economic importance of financial literacy: Theory and evidence," *Journal of Economic Literature*, 52, 5–44.

——— (2023): "The Importance of Financial Literacy: Opening a New Field," *Journal of Economic Perspectives*, 37, 137–154.

Oaxaca, R. (1973): "Male-female wage differentials in urban labor markets," *International Economic Review*, 14, 693–709.

Oaxaca, R. L. and M. Ransom (1994): "On discrimination and the decomposition of wage differentials," *Journal of Econometrics*, 61.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Statistics Canada (2020a): Canadian Perspectives Survey Series 5: Technology Use and Cyber Security during the Pandemic.

——— (2020b): Canadian Perspectives Survey Series 4: Information Sources Consulted During the Pandemic Public Use Microdata File, 2020.

Valdes, O. M., G. R. Mottola, J. T. Lin, and C. Bumcrot (2024): "The FINRA Foundation's National Financial Capability Study: Unpacking 12 years of data on U.S. financial capability," *Journal of Financial Literacy and Wellbeing*, 2, 79–90.

van Rooij, M., A. Lusardi, and R. Alessie (2011): "Financial literacy and stock market participation," *Journal of Financial Economics*, 101, 449–472.

7. Tables

Table 1: Weighted frequencies of respondents in each demographic group, by treatment

	Group A	Group B
Male	49.5	49.1
Female	50.5	50.9
Aged 18-34	27.7	28.8
Aged 35-54	31.9	32.6
Aged 55+	40.4	38.6
High school	41.5	43.3
College	31.9	28.8
University	26.6	27.9
British Columbia	14.1	13.8
Prairies	18.4	16.9
Ontario	38.9	39.2
Quebec	22.4	23.1
Atlantic	6.1	7.1
Low income	21.7	22.5
Medium income	27.2	27.8
High income	51.1	49.7
Unmarried	38.3	39.9
Married	61.7	60.1
Doesn't have kids	76.5	77.4
Has kids	23.5	22.6

Note: This table presents the weighted frequencies of respondents for each demographic category, separated by treatment group assignment (Group A versus Group B). Group A had the financial literacy questions at the beginning of the survey while Group B had the financial literacy questions at the end of the survey. The table provides a summary of the representation by gender, age, educational attainment, region, income level, marital status, and parental status.

Table 2: Percentage of respondents answering correctly, by demographic group

	Group A			Group B		
	Interest	Inflation	Risk	Interest	Inflation	Risk
Overall	90.5	69.3	70.3	82.4	64.1	55.2
Male	92.1	74.2	76.5	86.2	72.8	61.3
Female	88.9	64.5	64.1	78.8	55.7	49.3
Aged 18-34	86.3	51.4	53.5	79.1	47.1	47.6
Aged 35-54	91.1	69.9	70.4	80.6	58.0	53.1
Aged 55+	92.9	81.2	81.6	86.3	81.8	62.6
High school	85.8	59.1	60.5	74.5	52.8	43.1
College	92.9	70.7	74.1	84.5	67.2	56.4
University	94.9	83.7	80.8	92.4	78.3	72.8
British Columbia	93.5	68.6	72.9	90.9	75.7	62.8
Prairies	88.1	72.3	69.1	83.7	68.6	54.2
Ontario	91.4	71.5	72.1	82.6	64.0	56.0
Quebec	88.5	65.4	67.6	81.4	56.6	54.3
Atlantic	91.8	62.7	65.5	64.6	55.5	40.9
Low income	84.5	56.0	55.0	73.6	48.6	40.7
Medium income	89.7	68.9	69.5	85.4	64.0	56.2
High income	94.7	77.2	80.3	86.0	72.3	63.2
Unmarried	88.2	64.8	64.1	80.2	55.7	49.8
Married	91.9	72.2	74.1	83.8	69.6	58.7
Doesn't have kids	90.9	71.5	72.0	82.0	66.2	55.7
Has kids	89.2	62.2	64.6	83.7	56.7	53.4

Note: This table reports the weighted percentage of respondents in each demographic group who answered each of the three financial literacy questions (interest rate, inflation, risk diversification) correctly, separately for Group A (financial literacy questions at the beginning of the survey) and Group B (financial literacy questions at the end of the survey). Higher values indicate a greater share of correct responses within that subgroup.

Table 3: Percentage of respondents answering incorrectly, by demographic group

	Group A			Group B		
	Interest	Inflation	Risk	Interest	Inflation	Risk
Overall	6.2	22.8	6.7	9.2	20.8	7.4
Male	5.3	21.0	8.3	9.3	18.6	10.5
Female	7.2	24.5	5.1	9.2	22.7	4.5
Aged 18-34	9.6	35.8	16.5	11.3	32.5	12.4
Aged 35-54	5.6	21.3	3.7	10.7	24.4	7.2
Aged 55+	4.4	15.0	2.3	6.4	9.0	4.0
High school	9.0	29.4	8.4	13.2	25.5	8.9
College	4.8	23.2	5.4	7.7	18.0	7.2
University	3.7	11.9	5.5	4.5	16.2	5.4
British Columbia	4.6	24.0	2.9	2.8	11.6	3.1
Prairies	9.9	19.4	5.3	7.6	18.8	6.4
Ontario	5.6	20.9	6.0	10.8	22.5	10.0
Quebec	6.0	25.9	10.6	9.3	22.7	4.3
Atlantic	3.5	29.9	9.4	16.2	26.8	14.5
Low income	7.5	30.3	11.7	14.1	28.8	7.6
Medium income	7.0	23.5	6.5	8.2	22.6	7.7
High income	4.9	19.0	4.6	8.1	16.7	7.3
Unmarried	7.1	25.4	10.4	9.7	26.1	10.6
Married	5.7	21.1	4.4	8.9	17.2	5.3
Doesn't have kids	5.7	21.0	5.1	9.2	18.8	7.1
Has kids	8.0	28.4	11.7	9.3	27.5	8.5

Note: This table reports the weighted percentage of respondents in each demographic group who answered each of the three financial literacy questions incorrectly (interest rate, inflation, risk diversification), separately for Group A and Group B. Higher values indicate a greater share of incorrect responses within that subgroup.

Table 4: Percentage of respondents answering ”Don’t know,” by demographic group

	Group A			Group B		
	Interest	Inflation	Risk	Interest	Inflation	Risk
Overall	3.3	7.9	23.1	8.4	15.2	37.4
Male	2.7	4.8	15.2	4.6	8.6	28.1
Female	3.9	10.9	30.8	12.0	21.6	46.3
Aged 18-34	4.1	12.8	30.0	9.6	20.4	40.0
Aged 35-54	3.2	8.7	25.9	8.7	17.6	39.7
Aged 55+	2.7	3.9	16.1	7.3	9.2	33.4
High school	5.2	11.5	31.1	12.3	21.6	48.1
College	2.4	6.1	20.5	7.8	14.8	36.4
University	1.4	4.4	13.6	3.1	5.6	21.7
British Columbia	1.8	7.3	24.2	6.3	12.7	34.0
Prairies	2.0	8.2	25.6	8.7	12.6	39.5
Ontario	2.9	7.6	21.9	6.6	13.5	34.0
Quebec	5.5	8.7	21.8	9.3	20.7	41.4
Atlantic	4.6	7.4	25.1	19.2	17.6	44.6
Low income	8.0	13.7	33.3	12.3	22.5	51.7
Medium income	3.3	7.6	24.0	6.4	13.4	36.1
High income	0.4	3.8	15.2	5.9	10.9	29.6
Unmarried	4.7	9.7	25.5	10.1	18.2	39.6
Married	2.4	6.7	21.5	7.3	13.2	35.9
Doesn’t have kids	3.4	7.4	22.9	8.8	15.0	37.2
Has kids	2.8	9.4	23.7	7.0	15.8	38.1

Note: This table reports the weighted percentage of respondents in each demographic group who selected “Don’t know” for each of the three financial literacy questions (interest rate, inflation, risk diversification), separately for Group A and Group B. Higher values indicate a greater share of “Don’t know” responses within that subgroup.

Table 5: A/B test results: Correct, incorrect, and DK response rates

Question	Group A	Group B	Difference	With controls
Correct Answer Rate				
Interest rate	0.90	0.82	-0.08***	-0.08***
Inflation	0.69	0.64	-0.05**	-0.04**
Risk diversification	0.70	0.55	-0.15***	-0.15***
Incorrect Answer Rate				
Interest rate	0.06	0.09	0.03**	0.03**
Inflation	0.23	0.21	-0.02	-0.02
Risk diversification	0.07	0.07	0.01	0.01
DK Response Rate				
Interest rate	0.03	0.08	0.05***	0.04***
Inflation	0.08	0.15	0.07***	0.07***
Risk diversification	0.23	0.37	0.14***	0.14***

Note: Differences are tested using simple linear regression with weights and no controls for the first specification. Standard controls are added in the second: age, gender, region, education, income, marital status, and having kids [in the home?]. The significance levels are denoted as *** p <0.01, ** p<0.05, * p < 0.1.

Table 6: Oaxaca–Blinder results: Interest rate question

	Correct	Incorrect	Don't Know
Overall			
Group A	0.91***	0.06***	0.03***
	(0.01)	(0.01)	(0.00)
Group B	0.83***	0.10***	0.07***
	(0.01)	(0.01)	(0.01)
Difference	0.08***	-0.03***	-0.05***
	(0.02)	(0.01)	(0.01)
Endowments	0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
Coefficients	0.08***	-0.03**	-0.04***
	(0.02)	(0.01)	(0.01)
Endowments			
Sig. Covars	None	None	None
Coefficients			
Sig. Covars	Educ, Region, Constant	Region, Constant	Gender, Educ, Region

Note: This table reports the Oaxaca–Blinder decompositions of the treatment gap in the interest rate financial literacy question and in correct-response rates for each question between Group A and Group B. The "Endowments" component captures the part of the gap explained by differences in observed characteristics (age, gender, region, education, income, marital status, children at home), while the "Coefficients" component reflects differences in how these characteristics relate to outcomes across groups, including potential survey-design and fatigue effects; statistically significant covariates contributing to each component are listed in the last rows.

Table 7: Oaxaca–Blinder results: Inflation question

	Correct	Incorrect	Don't Know
Overall			
Group A	0.70***	0.23***	0.07***
	(0.02)	(0.01)	(0.01)
Group B	0.65***	0.21***	0.14***
	(0.02)	(0.01)	(0.01)
Difference	0.06**	0.02	-0.07***
	(0.02)	(0.02)	(0.02)
Endowments	0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.00)
Coefficients	0.04**	0.02	-0.07***
	(0.02)	(0.02)	(0.01)
Endowments			
Sig. Covars	None	None	None
Coefficients			
Sig. Covars	Region	Educ, Region	Educ

Note: This table reports the Oaxaca–Blinder decompositions of the treatment gap in the inflation financial literacy question and in correct-response rates for each question between Group A and Group B. The "Endowments" component captures the part of the gap explained by differences in observed characteristics (age, gender, region, education, income, marital status, children at home), while the "Coefficients" component reflects differences in how these characteristics relate to outcomes across groups, including potential survey-design and fatigue effects; statistically significant covariates contributing to each component are listed in the last rows.

Table 8: Oaxaca–Blinder results: Risk question

	Correct	Incorrect	Don't Know
Overall			
Group A	0.72***	0.07***	0.22***
	(0.01)	(0.01)	(0.01)
Group B	0.56***	0.07***	0.36***
	(0.02)	(0.01)	(0.02)
Difference	0.16***	-0.01	-0.15***
	(0.02)	(0.01)	(0.02)
Endowments	0.01	-0.00	-0.01
	(0.01)	(0.00)	(0.01)
Coefficients	0.15***	-0.01	-0.14***
	(0.02)	(0.01)	(0.02)
Endowments			
Sig. Covars	None	None	None
Coefficients			
Sig. Covars	Age, Educ	Age, Income, Region	Constant

Note: This table reports the Oaxaca–Blinder decompositions of the treatment gap in the risk financial literacy question and in correct-response rates for each question between Group A and Group B. The "Endowments" component captures the part of the gap explained by differences in observed characteristics (age, gender, region, education, income, marital status, having children at home), while the "Coefficients" component reflects differences in how these characteristics relate to outcomes across groups, including potential survey-design and fatigue effects; statistically significant covariates contributing to each component are listed in the last rows.

Table 9: Regression of correct, incorrect, and DK responses on treatment and gender

	Correct		*Incorrect*		*DK*	
	#	> 0	#	> 0	#	> 0
Treatment	-0.24***	-0.03**	0.04	0.05	0.19***	0.15***
	(0.06)	(0.02)	(0.04)	(0.03)	(0.04)	(0.03)
Female	-0.22***	0.00	0.02	0.06*	0.20***	0.16***
	(0.05)	(0.01)	(0.04)	(0.03)	(0.04)	(0.03)
Treatment × female	-0.06	-0.04	-0.06	-0.06	0.12*	0.03
	(0.08)	(0.02)	(0.06)	(0.04)	(0.07)	(0.04)
Controls?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,502	3,502	3,502	3,502	3,502	3,502

Note: The total number of correct responses is denoted as #. This is similar for incorrect and DK responses. The indicator “>0” takes the value of 1 if a respondent answers any FL question correctly/incorrectly/DK. Treatment takes on a value of 1 if an individual is in Group B. Covariates include the standard set: age, gender, region, education, income, marital status, and having kids [in the home]. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

Table 10: Regression: Number of correct, incorrect, and DK responses on treatment and age group

	Correct		*Incorrect*		*DK*	
	#	> 0	#	> 0	#	> 0
Treatment	-0.20**	-0.06**	-0.04	-0.03	0.24***	0.13***
	(0.08)	(0.02)	(0.06)	(0.04)	(0.06)	(0.04)
Age: 35-54	0.34***	0.03*	-0.29***	-0.18***	-0.06	-0.03
	(0.07)	(0.02)	(0.06)	(0.04)	(0.05)	(0.04)
Age: 55+	0.67***	0.06***	-0.41***	-0.27***	-0.27***	-0.18***
	(0.07)	(0.02)	(0.05)	(0.04)	(0.05)	(0.03)
Treatment × Age: 35-54	-0.17	-0.01	0.14*	0.10*	0.03	0.03
	(0.11)	(0.03)	(0.08)	(0.06)	(0.09)	(0.05)
Treatment × Age: 55+	-0.03	0.03	0.03	0.02	0.01	0.05
	(0.10)	(0.03)	(0.07)	(0.05)	(0.08)	(0.05)
Controls?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,502	3,502	3,502	3,502	3,502	3,502

Note: The total number of correct responses is denoted as #. This is similar for incorrect and DK responses. The indicator “>0” takes the value of 1 if a respondent answers any FL question correctly/incorrectly/DK. Treatment takes on a value of 1 if an individual is in Group B. Covariates include the standard set: age, gender, region, education, income, marital status, and having kids [in the home]. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

Table 11: Regression: Number of correct, incorrect, and DK responses on treatment and education group

	Correct		*Incorrect*		*DK*	
	#	> 0	#	> 0	#	> 0
Treatment	-0.33***	-0.08***	0.01	0.01	0.32***	0.19***
	(0.08)	(0.02)	(0.06)	(0.04)	(0.06)	(0.04)
College	0.23***	0.02*	-0.08	-0.05	-0.15***	-0.11***
	(0.06)	(0.01)	(0.05)	(0.04)	(0.05)	(0.03)
University	0.46***	0.03**	-0.24***	-0.19***	-0.22***	-0.17***
	(0.06)	(0.01)	(0.05)	(0.04)	(0.05)	(0.03)
Treatment × College	0.06	0.02	-0.02	-0.02	-0.04	-0.01
	(0.10)	(0.03)	(0.07)	(0.05)	(0.08)	(0.05)
Treatment × University	0.17*	0.07***	0.03	0.05	-0.20***	-0.11**
	(0.09)	(0.02)	(0.06)	(0.05)	(0.07)	(0.05)
Controls?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,502	3,502	3,502	3,502	3,502	3,502

Note: The total number of correct responses is denoted as #. This is similar for incorrect and DK responses. The indicator “>0” takes the value of 1 if a respondent answers any FL question correctly/incorrectly/DK. Treatment takes on a value of 1 if an individual is in Group B. Covariates include the standard set: age, gender, region, education, income, marital status, and having kids [in the home]. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

Table 12: Oaxaca–Blinder robustness check: Cash holdings

	Cash on hand		Other cash	
	$	Log	$	Log
Overall				
Group A	123.02***	3.42***	172.55***	1.31***
	(6.90)	(0.07)	(22.90)	(0.08)
Group B	118.39***	3.33***	157.06***	1.28***
	(7.22)	(0.07)	(17.79)	(0.08)
Difference	4.63	0.09	15.49	0.03
	(9.99)	(0.10)	(29.00)	(0.11)
Endowments	1.44	0.03	6.23	0.02
	(1.76)	(0.03)	(4.30)	(0.02)
Coefficients	3.20	0.06	9.26	0.01
	(9.85)	(0.10)	(28.10)	(0.11)
Endowments				
Sig. Covars	None	None	None	None
Coefficients				
Sig. Covars	None	Educ	Income, Region	Income, Region

Note: The Table reports Oaxaca–Blinder decompositions of the difference in cash holdings between Group A and Group B, using two measures: dollar levels and log amounts of cash on hand and other cash. "Endowments" captures the differences in observable characteristics (e.g., education, income, region), while "Coefficients" captures differences in how these characteristics relate to cash holdings across groups; rows at the bottom list which covariates significantly contribute to each component. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

Table 13: Robustness Oaxaca–Blinder check: Interest question, CPSS questions included

	Correct	Incorrect	Don't know
Overall			
Group A	0.91***	0.06***	0.03***
	(0.01)	(0.01)	(0.00)
Group B	0.83***	0.10***	0.07***
	(0.01)	(0.01)	(0.01)
Difference	0.08***	-0.03***	-0.05***
	(0.02)	(0.01)	(0.01)
Endowments	0.01*	-0.00	-0.00*
	(0.00)	(0.00)	(0.00)
Coefficients	0.07***	-0.03**	-0.04***
	(0.02)	(0.01)	(0.01)
Endowments			
Sig. Covars	None	None	None
Sum of residuals	0.00**	-0.00*	-0.00*
	(0.00)	(0.00)	(0.00)
Coefficients			
Sig. Covars	Educ, Region, Constant	Region, Constant	Gender, Educ, Region
Sum of residuals	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)

Note: This table reports the Oaxaca–Blinder decompositions of the treatment gap in the interest financial literacy question and in correct-response rates for each question between Group A and Group B. The "Endowments" component captures the part of the gap explained by differences in observed characteristics (age, gender, region, education, income, marital status, having children at home), while the "Coefficients" component reflects differences in how these characteristics relate to outcomes across groups, including potential survey-design and fatigue effects; statistically significant covariates contributing to each component are listed in the last rows. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

Table 14: Robustness Oaxaca–Blinder check: Inflation question, CPSS questions included

	Correct	Incorrect	Don't know
Overall			
Group A	0.70***	0.23***	0.07***
	(0.02)	(0.01)	(0.01)
Group B	0.65***	0.21***	0.14***
	(0.02)	(0.01)	(0.01)
Difference	0.06**	0.02	-0.07***
	(0.02)	(0.02)	(0.02)
Endowments	0.02*	-0.01	-0.01*
	(0.01)	(0.01)	(0.00)
Coefficients	0.04*	0.03	-0.07***
	(0.02)	(0.02)	(0.01)
Endowments			
Sig. Covars	None	None	None
Sum of residuals	0.00*	-0.00	-0.00*
	(0.00)	(0.00)	(0.00)
Coefficients			
Sig. Covars	None	Educ, Region	Educ, Region
Sum of residuals	-0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)

Note: This table reports the Oaxaca–Blinder decompositions of the treatment gap in the inflation financial literacy question and in correct-response rates for each question between Group A and Group B. The "Endowments" component captures the part of the gap explained by differences in observed characteristics (age, gender, region, education, income, marital status, having children at home), while the "Coefficients" component reflects differences in how these characteristics relate to outcomes across groups, including potential survey-design and fatigue effects; statistically significant covariates contributing to each component are listed in the last rows. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

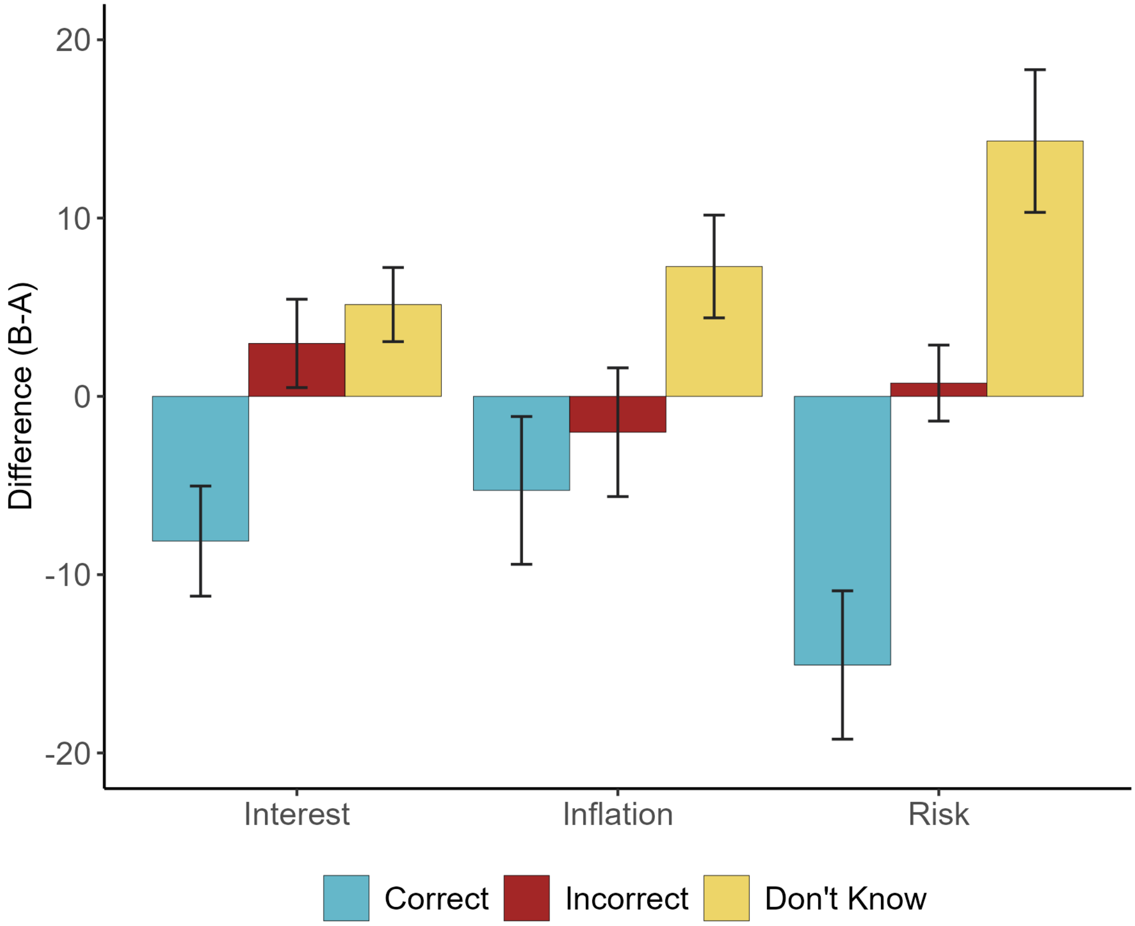
Table 15: Robustness Oaxaca–Blinder check: Risk question, CPSS questions included

	Correct	Incorrect	Don't know
Overall			
Group A	0.72***	0.07***	0.22***
	(0.01)	(0.01)	(0.01)
Group B	0.56***	0.07***	0.36***
	(0.02)	(0.01)	(0.02)
Difference	0.16***	-0.01	-0.15***
	(0.02)	(0.01)	(0.02)
Endowments	0.01*	-0.00	-0.01
	(0.01)	(0.00)	(0.01)
Coefficients	0.14***	-0.00	-0.14***
	(0.02)	(0.01)	(0.02)
Endowments			
Sig. Covars	None	None	None
Sum of residuals	0.01**	-0.00*	-0.00*
	(0.00)	(0.00)	(0.00)
Coefficients			
Sig. Covars	Age, Educ	Age, Income, Region	None
Sum of residuals	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)

Note: This table reports the Oaxaca–Blinder decompositions of the treatment gap in the risk financial literacy question and in correct-response rates for each question between Group A and Group B. The "Endowments" component captures the part of the gap explained by differences in observed characteristics (age, gender, region, education, income, marital status, having children at home), while the "Coefficients" component reflects differences in how these characteristics relate to outcomes across groups, including potential survey-design and fatigue effects; statistically significant covariates contributing to each component are listed in the last rows. Standard errors are in parentheses with *** p<0.01, ** p<0.05, and * p<0.1, respectively.

8. Figures

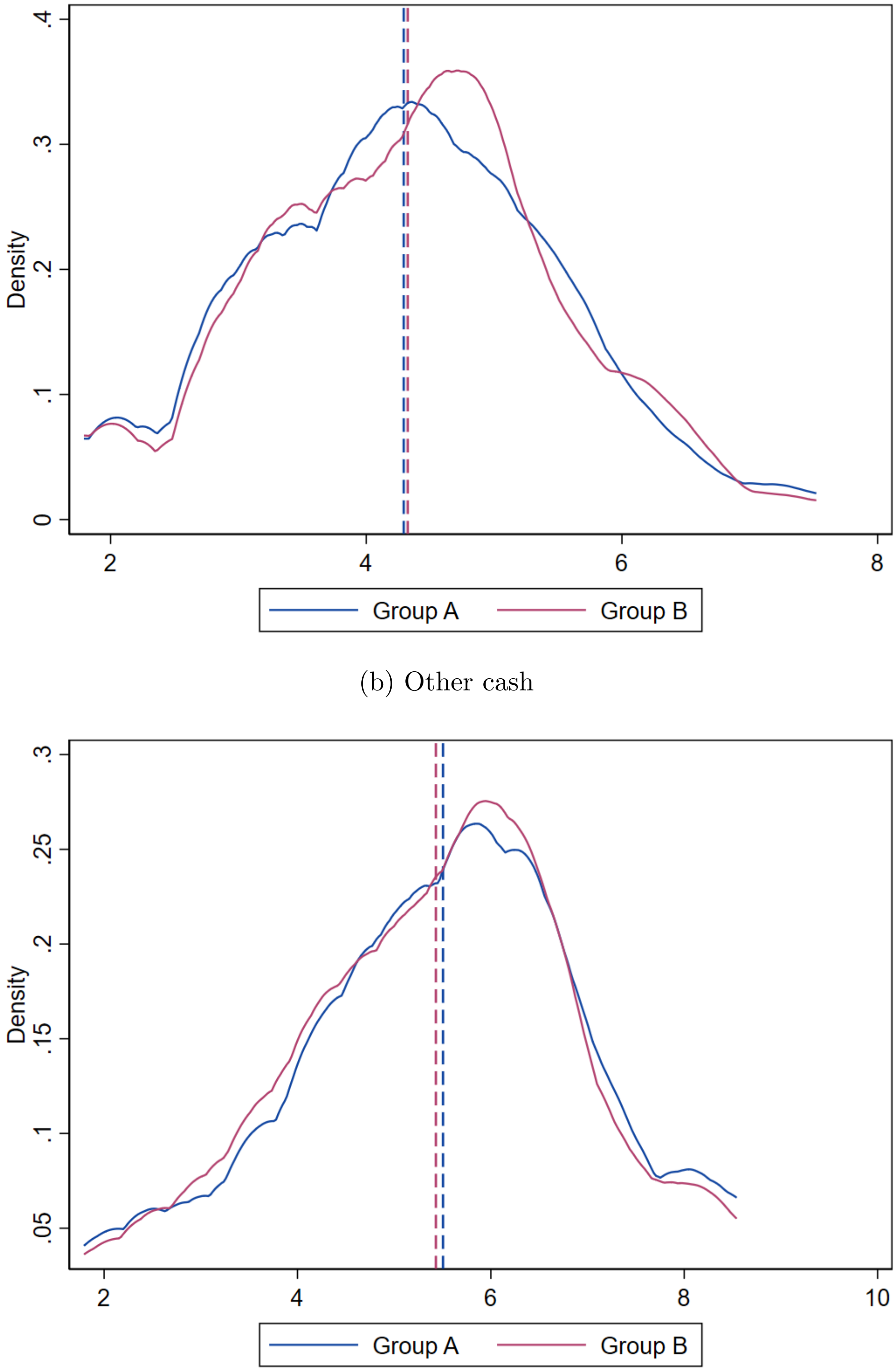
Figure 1: Difference in response shares in Group B, relative to Group A, in percentage points



Note: This figure shows the difference in shares between Group B and Group A. Survey weights are applied. Standard errors are calculated using simple linear regression of response type on treatment (indicator for being in Group B), with no controls.

Figure 2: Kernel density of log cash holdings

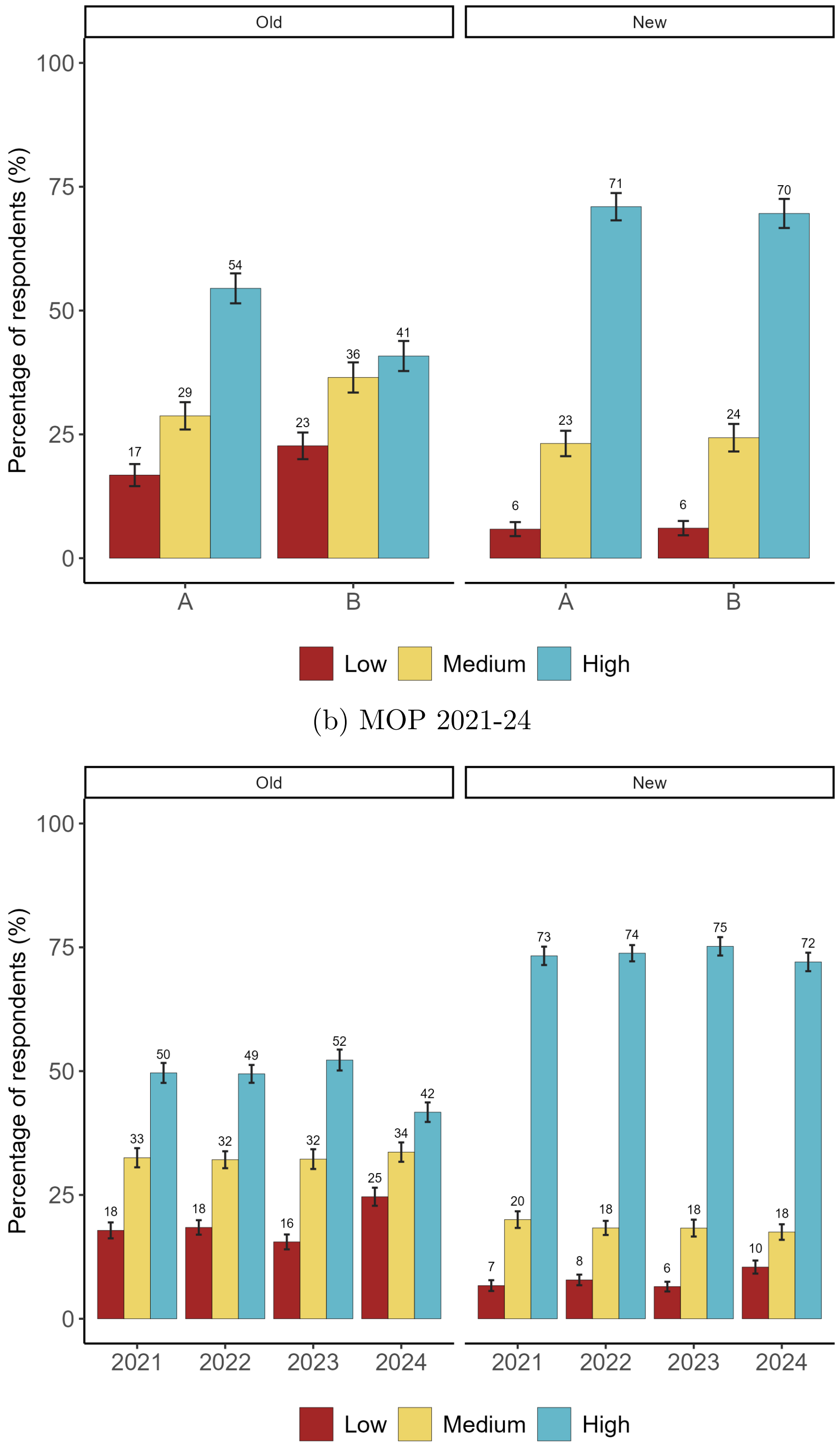
(a) Cash on hand



(b) Other cash

Note: Both "Cash on hand" and "Other cash" distributions overlap for groups A and B.

Figure 3: Two measures of financial literacy (old vs. new)



(a) CAS (A/B)

(b) MOP 2021-24

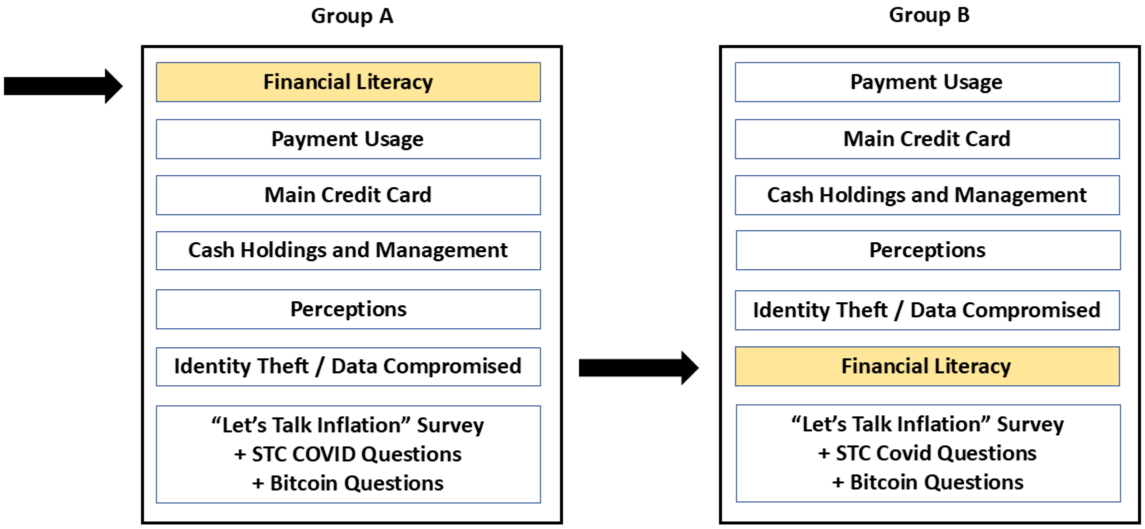
Note: The "old" index refers to one used in previous Methods-of-Payment Survey reports, which uses correct, incorrect, and DK responses, see Henry et al. (2024a). The "new" index is one where only the number of incorrect responses matters: No incorrect responses nets a "high" financial literacy score, 1 incorrect response nets a "medium" score, and 2-3 incorrect responses nets a "low" score.

A. Appendix

A.1. Survey layout

The November 2020 Cash Alternative Survey (CAS) survey questionnaire is formatted as follows.

Figure A-1: Layout of the November 2020 CAS



The following are the "Big Three" financial literacy questions, in order (**?**).

Table A-1: "Big Three" financial literacy questions and correct responses

Question	Response options
Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have left in the account if you left the money to grow?	**More than $102** Exactly $102 Less than $102 Don't know
Imagine the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with this money in this account?	More than today Exactly the same **Less than today** Don't know
Please tell me whether or not this statement is true or false: Buying a single company's stock usually provides a safer return than a mutual fund of stocks.	True **False** Don't know

A.2. Chi-square test

The chi-square test of independence tests the following distributional hypotheses for correct, incorrect, and "Don't know" (DK) responses:

- **Null hypothesis** (H_0): Response categories of groups A and B follow an identical distribution.

- **Alternative hypothesis** (H_1): There is a different distribution of response categories between the groups.

Contingency Table Structure:

	Correct	Incorrect	Don't know	Total
Group A	a_1	a_2	a_3	n_A
Group B	b_1	b_2	b_3	n_B
Total	t_1	t_2	t_3	n

where a_i and b_i denote the counts in each response category for Group A and Group B, n_A and n_B are the group totals, t_j are the column totals, and n is the overall sample size.

The chi-square test of independence for a 2 × 3 table computes the probability of observing the given table under the null hypothesis, conditional on the observed row and column totals. The p-value is calculated with the computed χ^2 statistic and corresponding degrees of freedom, which is $= (N_{row} - 1) \times (N_{col} - 1) = 2$ for a 2×3 contingency table. The chi-square statistic is computed with the following formula:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(c_{ij} - E_{ij})^2}{n},$$

where c_{ij} is the count in a given cell (row = i, col=j), E_{ij} is the expected number of responses in a cell under the null hypothesis, and n is the total number of observations.

If the computed p-value is less than the significance threshold (e.g., $\alpha = 0.05$), we reject the null hypothesis and conclude that the distribution of responses differs significantly between the two groups. The chi-square test is preferred in this context (as opposed to Fisher's exact test) because our sample size is sufficiently large, with high counts in each cell.

The contingency tables for each question are as follows:

Table A-2: Interest question

	Correct	Incorrect	DK	Total
Group A	1,755	121	70	1,946
Group B	1,625	180	141	1,946
Total	3,380	301	211	3,892

Chi-square p-value: 0.00

df = 2

Table A-3: Inflation question

	Correct	Incorrect	DK	Total
Group A	1,343	452	151	1,946
Group B	1,279	412	255	1,946
Total	2,622	864	406	3,892

Chi-square p-value: 0.00

df = 2

Table A-4: Risk question

	Correct	Incorrect	DK	Total
Group A	1,317	172	457	1,946
Group B	1,108	158	680	1,946
Total	2,425	330	1,137	3,892

Chi-square p-value: 0.00

df = 2

A.3. Robustness: Inverse Probability Weighting

While the random assignment of financial literacy (FL) question placement in our experiment supports causal inference, concerns remain regarding potential differences in survey uptake or unmeasured confounding between Groups A and B. To address these concerns and further strengthen the robustness of our results, we implement an inverse probability weighting (IPW) procedure, which allows us to reweight observations to account for observed covariate imbalances and estimate average treatment effects in a manner analogous to a randomized experiment (Rosenbaum and Rubin, 1983; Imai et al., 2008).

The IPW approach begins by estimating each respondent's probability of being assigned to the treatment group (Group B) conditional on observed characteristics. Specifically, we fit a logistic regression model:

$$P(T_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}, \tag{10}$$

where T_i is an indicator for treatment assignment (1 if Group B, 0 if Group A), and X_i is a vector of observed covariates for respondent i (including age, gender, education, income, marital status, parental status, region, and their interactions).

For each individual, we then compute the inverse probability weight as follows:

$$w_i = \begin{cases} \frac{1}{\hat{P}(T_i=1|X_i)} & \text{if } T_i = 1 \\ \frac{1}{1-\hat{P}(T_i=1|X_i)} & \text{if } T_i = 0 \end{cases}, \tag{11}$$

where $\hat{P}(T_i = 1|X_i)$ is the estimated propensity score for individual i.

These weights are then applied in the estimation of treatment effects, such that the weighted sample more closely resembles a pseudo-population in which treatment assignment is independent of observed covariates. The average treatment effect (ATE) is estimated as:

$$ATE_{IPW} = \frac{\sum_i w_i T_i Y_i}{\sum_i w_i T_i} - \frac{\sum_i w_i (1 - T_i) Y_i}{\sum_i w_i (1 - T_i)}, \tag{12}$$

where Y_i is the outcome of interest (the number of correct, incorrect, and "Don't know" responses) for respondent i.

To assess the robustness of our main findings, we conduct IPW analyses for the key outcomes (the number of correct, incorrect, and "Don't know" responses), but we also conduct subgroup analyses by gender, age, education, and Bitcoin ownership.

Finally, we perform diagnostic checks to ensure adequate overlap in propensity scores and assess the balance of covariates after weighting. By incorporating IPW alongside other robustness checks, we provide a comprehensive assessment of the causal impact of financial literacy (FL) question placement on survey responses, while explicitly addressing concerns about both observed and potential unobserved confounding.

A.4. Inverse Probability Weighting

The main inverse probability weighting (IPW) results in Table A-5 provide further insights into the effects of treatment on financial literacy outcomes by indirectly accounting for some sample selection bias. Unlike in the Oaxaca–Blinder (OB) decomposition, the matching methods first estimate a propensity score for treatment, based on a set of covariates, and then estimate the treatment effect using matched pairings of variables. This analysis does not incorporate the use of weights. Results differ slightly from the gaps estimated in the Oaxaca–Blinder decomposition.

The PSM results related to the correct responses, as shown in Table A-5, indicate a consistent, negative average treatment effect (ATE) for correct responses across all three questions (Q1, Q2, Q3). Specifically, the ATE values are -0.07, -0.03, and -0.11, all statistically significant at the $p < 0.01$ level.[2] This suggests that respondents in Group A performed significantly better than those in Group B. These findings align with the OB decomposition results, which also show that placement significantly correlates with performance on financial literacy questions. The OB analysis highlighted that question placement leads to cognitive effects such as survey fatigue or convenience, which may explain why respondents placed later in the survey tend to provide fewer correct answers.

For DK responses (also **Table A-5** and **Appendix Tab??)**, the PSM results reveal significant positive ATEs for DK responses across all questions: +0.04 for Question 1, +0.05 for Question 2, and +0.12 for Question 3. This indicates that respondents in Group B were more likely to select DK compared with those in Group A. This reinforces the OB findings, which suggest that question placement significantly impacts respondents' likelihood of expressing uncertainty. The OB analysis notes that uncertainty could be influenced by cognitive load or fatigue associated with answering questions later in a survey. The PSM results thus corroborate these findings by demonstrating that later question placement increases DK responses.

In terms of incorrect responses (Table A-6), the PSM results show negligible effects for Questions 2 and 3 (-0.02 and -0.01,respectively), with only Question 1 (+0.03) being statistically significant. This suggests that some respondents may be prone to answering incorrectly when FL questions are placed later in the survey. This inconsistency can be linked back to the Oaxaca–Blinder findings, which indicate that demographic factors such as gender might influence performance differently, based on question placement. The Oaxaca–Blinder results show that women were more likely to provide incorrect answers when the [financial literacy?] questions were placed later in the survey. Thus, the PSM results reinforce the notion that demographic characteristics can mediate the treatment effects observed in financial literacy assessments.

Put together, these results extend the work of Huynh and Voia (2024) by demonstrating a causal relationship between question placement and response patterns. The findings also support the research of Lusardi and Mitchell (2014) on the importance of survey design in

[2] The ATT estimates are nearly identical to the ATE (Appendix: Table **??**).

Table A-5: Propensity score matching results: Average treatment effect (ATE)

	Correct		DK	
	(1)	(2)	(3)	(4)
Question	AGR	All	AGR	All
Interest rate	-0.07***	-0.07***	0.04***	0.04***
	(0.01)	(0.01)	(0.01)	(0.01)
Inflation	-0.03*	-0.03*	0.05***	0.05***
	(0.01)	(0.02)	(0.01)	(0.01)
Risk diversification	-0.11***	-0.11***	0.12***	0.12***
	(0.02)	(0.02)	(0.01)	(0.01)
Observations	3,881	3,502	3,881	3,502

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Matching variables in "AGR" include age, gender, and region. Matching variables in "All" include the full standard set of covariates: age, gender, education, income, marital status, and having kids at home.

Table A-6: Propensity score matching results: Average treatment effect (ATE)

	Incorrect	
	(1)	(2)
Question	AGR	All
Interest rate	0.03***	0.03***
	(0.01)	(0.01)
Inflation	-0.02	-0.02
	(0.01)	(0.01)
Risk diversification	-0.01	-0.01
	(0.01)	(0.01)
Observations	3,881	3,502

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Matching variables in "AGR" include age, gender, and region. Matching variables in "All" include the full standard set of covariates: age, gender, education, income, marital status, and having kids at home.

financial literacy assessment.

A.5. Heterogeneity: Additional analysis

In this section, we provide additional heterogeneity analysis.

A.5.1. Gender

Tables A-7 (logit) and A-8 (ordered logit), together with Figure A-2, present the results of the heterogeneity analysis by gender. The data indicate that both male and female respondents are more likely to select "Don't know" (DK) when the financial literacy questions are positioned at the end of the survey (the treatment group, Group B), compared with when they are placed at the beginning (Group A). This increase in DK responses is accompanied by a corresponding decrease in correct answers, while the proportion of incorrect responses remains relatively stable across treatment groups.

The magnitude of the treatment effect on DK responses is higher for female respondents. Specifically, Table A-8 shows that the probability of a DK response for females rises from Group A to Group B, exceeding the increase observed among males. Table A-9 further quantifies these differences, indicating that the shift from correct to DK responses is more pronounced for females. This pattern is consistent with prior literature documenting gender differences in financial literacy and response confidence, where women are generally more likely to express uncertainty in survey settings; see Lusardi and Mitchell (2008).

Figure A-2 presents the estimated interaction effects between gender and treatment assignment on the probability of responding DK to any of the three financial literacy questions. These effects are computed following the approach of Ai et al. (2004) (using the `inteff` command in STATA). The figure displays the marginal effect of being female (relative to male) on the probability of a DK response, conditional on the treatment group. In Figure A-2, the vertical axis represents the change in the probability of a DK response due to the interaction between being female and being assigned to the treatment group (Group B, where the financial literacy questions are placed at the end of the survey). The horizontal axis represents the predicted probability of answering DK.

The results from Figure A-2 indicate that the presence of both positive and negative z-statistics, and the fact that the z-statistic crosses zero, highlights the heterogeneity in how the interaction between gender and treatment affects the likelihood of a DK response. This pattern suggests that the effect is not homogeneous across the sample; for some respondents (females with a low probability of answering DK), the placement of the financial literacy questions at the end of the survey increases the gender gap in DK responses, while for others (females with a higher probability of answering DK), it may reduce it or have no significant effect. The results reveal these nuanced, non-linear interaction effects that would be not visible by simply reporting the coefficient from a standard logit model.

Overall, the findings suggest that survey fatigue, induced by later question placement, affects both females and males but with a higher margin for the female respondents. The larger increase in DK responses among women may reflect lower confidence or greater susceptibility to fatigue effects, as has been observed in previous studies.

Table A-7: Results from logistic regression, female x treatment

			Correct	Incorrect	Don't know
			(>0)	(>0)	(>0)
$P(Y=1)$	Male	A	0.97	0.26	0.17
		B	0.93	0.31	0.32
	Female	A	0.96	0.32	0.33
		B	0.90	0.30	0.50
$\Delta P(Y=1)$	Male		-0.04**	0.05	0.15***
			(0.02)	(0.03)	(0.03)
	Female		-0.07***	-0.02	0.17***
			(0.02)	(0.03)	(0.03)

Note: *** p <0.01, ** p<0.05, * p < 0.1. $P(Y=1)$ refers to the estimated probability of each group having outcome Y (>0 correct, incorrect, or DK responses). $\Delta P(Y=1)$ corresponds to the treatment effect, estimated separately by gender. Logistic regression specification includes an interaction term between gender (female indicator) and treatment.

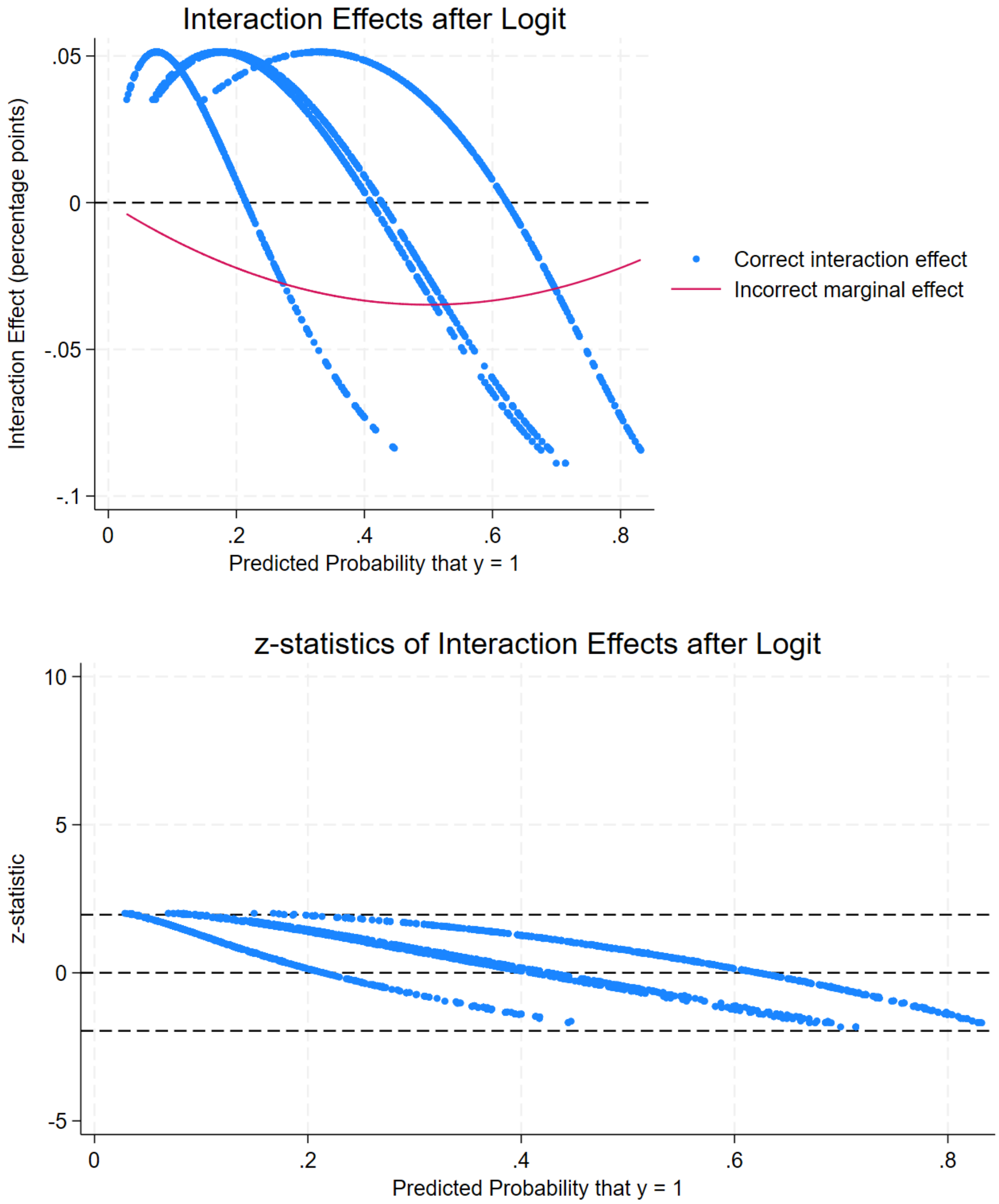
A.5.2. Age

Table A-10 presents the results from the logistic regression analysis of the interaction between age group and treatment assignment on the probability of providing a DK response to the financial literacy questions. The table indicates that treatment effects, that is, the increase in DK responses when the financial literacy questions are placed at the end of the survey (Group B) are present for all age groups, but their magnitude and statistical significance vary, being higher for the age group 55+.

Figures A-3 and A-4 further explore these interaction effects, using Ai et al. (2004), for the 35–54 and 55+ age groups relative to those aged 18–34. In Figure A-3, which focuses on respondents aged 35–54, the interaction effect of age and treatment on the probability of a DK response is generally positive across the sample. However, the corresponding z-statistic remains very close to zero throughout, indicating that the estimated interaction effect is not statistically significant for this age group relative to the 18–34 age group, which means they are mostly increasing their DK answers when moving from Group A to Group B almost at the same rate.

In contrast, Figure A-4, which examines respondents aged 55 and older, reveals a more nuanced pattern. For this group, the interaction effect relative to the youngest cohort can be positive, particularly at higher predicted probabilities of a DK response. This means that, for some older respondents, specifically those who are already more likely to select

Figure A-2: Interaction effects: Gender, any "Don't know" (DK) responses



Note: These plots show the estimated interaction effects by probability of responding DK to any of the three questions, as specified by Ai et al. (2004). The command *inteff* is used.

Table A-8: Results from ordered logistic regression, female x treatment

		Number correct				Number incorrect			
		0	1	2	3	0	1	2	3
$P(Y=1)$									
Male	A	0.03	0.11	0.24	0.62	0.73	0.22	0.05	0.01
	B	0.06	0.18	0.28	0.48	0.69	0.25	0.06	0.01
Female	A	0.06	0.17	0.28	0.48	0.69	0.25	0.06	0.01
	B	0.10	0.24	0.30	0.36	0.70	0.24	0.05	0.01
$\Delta P(Y=1)$									
Male		0.03***	0.06***	0.05***	-0.14***	-0.04	0.03	0.01	0.00
		(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.02)	(0.01)	(0.00)
Female		0.04***	0.07***	0.02***	-0.13***	0.01	-0.01	0.00	0.00
		(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.02)	(0.01)	(0.00)

		Number DK			
		0	1	2	3
$P(Y=1)$					
Male	A	0.83	0.13	0.02	0.01
	B	0.68	0.24	0.06	0.02
Female	A	0.67	0.25	0.06	0.02
	B	0.49	0.34	0.11	0.05
$\Delta P(Y=1)$					
Male		-0.15***	0.10***	0.03***	0.01***
		(0.03)	(0.02)	(0.01)	(0.00)
Female		-0.18***	0.10***	0.05***	0.03***
		(0.03)	(0.02)	(0.01)	(0.01)

Table A-9: Interaction ordered logistic regression: Gender x treatment, standard covariates (full table)

	(1) Number correct	(2) Number incorrect	(3) Number DK
Group B	-0.67***	0.23	0.92***
	(0.14)	(0.16)	(0.17)
Female	-0.65***	0.22	0.98***
	(0.14)	(0.15)	(0.17)
Group B × Female	0.03	-0.30	-0.08
	(0.19)	(0.22)	(0.22)
Covariates	All	All	All
/cut1	-2.34***	0.16	0.56**
	(0.21)	(0.23)	(0.23)
/cut2	-0.57***	2.17***	2.39***
	(0.20)	(0.24)	(0.23)
/cut3	0.94***	4.15***	3.73***
	(0.21)	(0.31)	(0.27)
Observations	3502	3502	3502

Note: The Stata command `ologit` is used to generate these tables. "/cut" corresponds to cutoff points of a latent continuous variable, $Num*$, which determine the observed count variable (for example, $NumDK$).

DK, the placement of the financial literacy questions at the end of the survey is associated with a greater probability of a DK response. However, this effect is not uniform across the entire age 55+ group and appears to be concentrated among those with higher baseline DK propensities.

Table A-10: Results from logistic regression, age x treatment

			Correct (> 0)	Incorrect (> 0)	Don't know (> 0)
$P(Y = 1)$	18-34	A	0.93	0.45	0.33
		B	0.88	0.42	0.47
	35-54	A	0.96	0.27	0.30
		B	0.88	0.35	0.46
	55+	A	0.99	0.19	0.16
		B	0.96	0.18	0.33
$\Delta P(Y = 1)$	18-34		-0.06**	-0.03	0.13***
			(0.02)	(0.04)	(0.04)
	35-54		-0.08***	0.08*	0.16***
			(0.02)	(0.04)	(0.04)
	55+		-0.03**	0.00	0.17***
			(0.01)	(0.03)	(0.03)

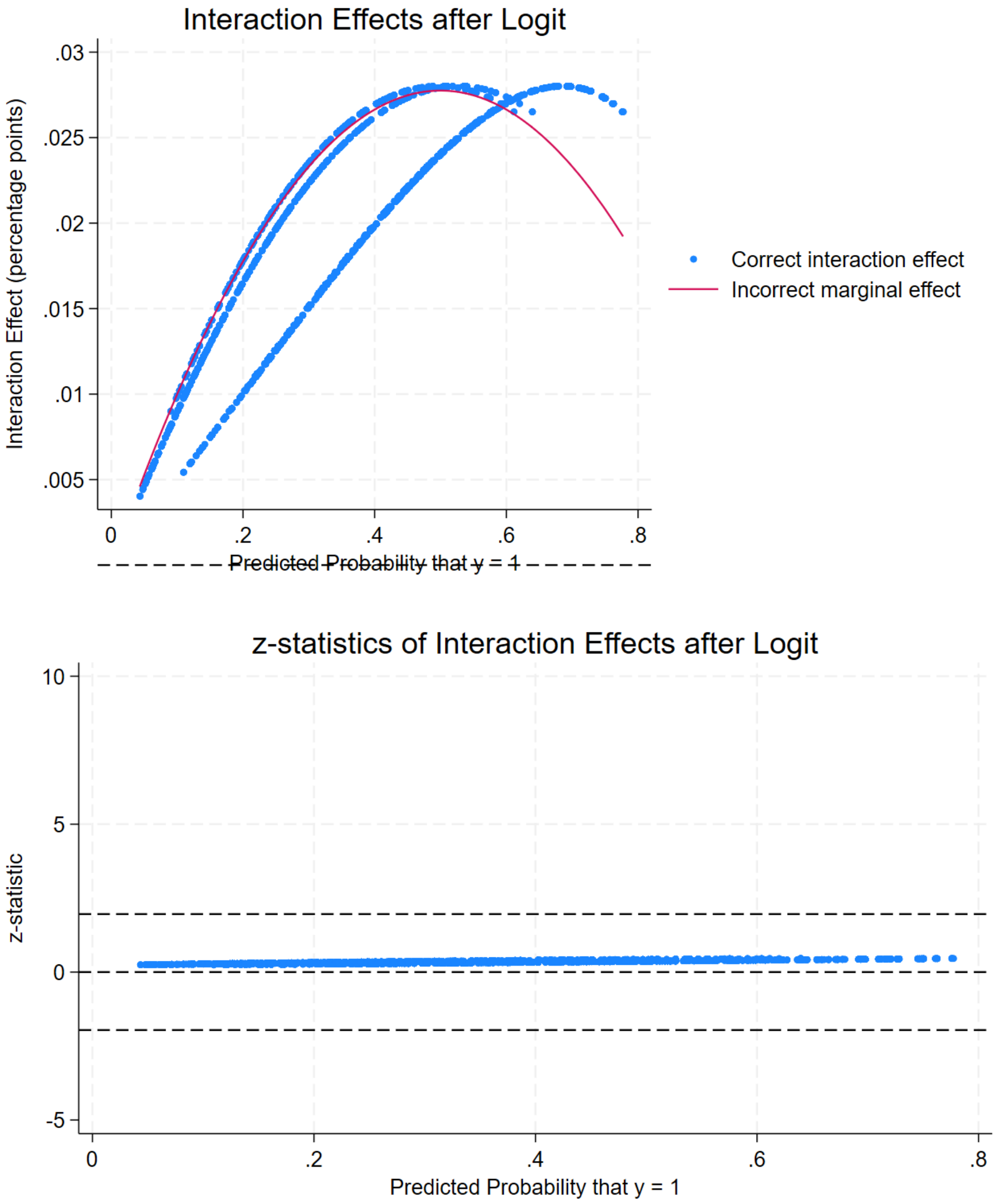
Note: *** p <0.01, ** p<0.05, * p < 0.1. $P(Y = 1)$ refers to the estimated probability of each group having outcome Y (> 0 correct, incorrect, or dk responses). $\Delta P(Y = 1)$ corresponds to the treatment effect, estimated separately by age category. Logistic regression specification includes an interaction term between age (age group indicator) and treatment.

A.5.3. Education

Table A-11 reports the results from the logistic regression examining the interaction between education level and treatment assignment on the probability of providing a DK response to the financial literacy questions. The table indicates that the treatment effect, the increase in DK responses when the financial literacy questions are placed at the end of the survey (Group B), is most pronounced among respondents with lower educational attainment (high school or college), while the effect is smaller for those with a university degree.

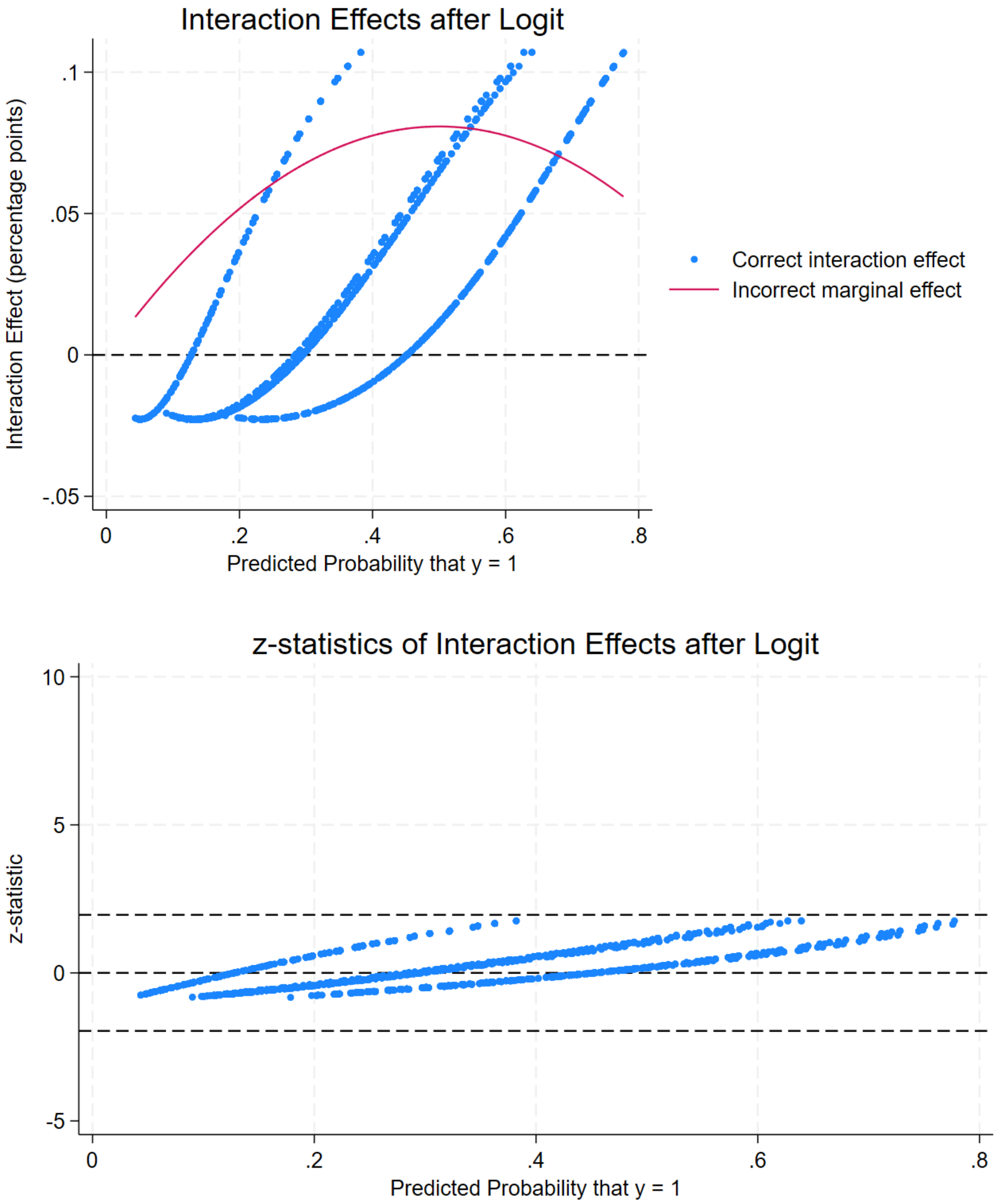
Figures A-5 and A-6 provide further insight into these interaction effects using Ai et al. (2004). Figure A-5 focuses on respondents with a college education (relative to high school), while Figure A-6 examines those with a university degree (relative to high school). In Figure A-5, the interaction effect of education and treatment on the probability of a DK response is generally positive, indicating that for college-educated respondents, being assigned to the treatment group increases the likelihood of selecting DK. However, the z-statistic for this interaction effect is close to zero across most of the distribution, suggesting that while

Figure A-3: Interaction effects: Age (35-54), any "Don't know" (DK) responses



Note: Age 18-34 is used as the base category in this case. These plots show the estimated interaction effects by probability of responding DK to any of the three questions, as specified by Ai et al. (2004). The command *inteff* is used.

Figure A-4: Interaction effects: Age (55+), any “Don’t know” (DK) responses



Note: Age 18-34 is used as the base category in this case. These plots show the estimated interaction effects by probability of responding DK to any of the three questions, as specified by Ai et al. (2004). The command *inteff* is used.

Table A-11: Results from ordered logistic regression, age x treatment

		Number correct				Number incorrect			
		0	1	2	3	0	1	2	3
$P(Y=1)$									
$18-34$	A	0.09	0.24	0.31	0.36	0.54	0.35	0.09	0.02
	B	0.13	0.29	0.31	0.28	0.56	0.34	0.09	0.02
$35-54$	A	0.04	0.15	0.28	0.53	0.73	0.22	0.04	0.01
	B	0.09	0.24	0.31	0.36	0.66	0.28	0.06	0.01
55+	A	0.02	0.08	0.20	0.71	0.82	0.15	0.03 0.00	
	B	0.04	0.13	0.26	0.57	0.82	0.15	0.03	0.00
$\Delta P(Y=1)$									
$18-34$		0.04**	0.05***	-0.01	-0.08***	0.02	-0.01	-0.01	0.00
		(0.02)	(0.02)	(0.00)	(0.03)	(0.04)	(0.03)	(0.01)	(0.00)
$35-54$		0.04***	0.09***	0.03***	-0.17***	-0.07*	0.05*	0.02* 0.00*	
		(0.01)	(0.02)	(0.01)	(0.03)	(0.04)	(0.03)	(0.01)	(0.00)
55+		0.02***	0.06***	0.07***	-0.14***	0.01	0.00	0.00	0.00
		(0.00)	(0.01)	(0.02)	(0.03)	(0.03)	(0.02)	(0.00)	(0.00)

		Number DK			
		0	1	2	3
$P(Y=1)$					
$18-34$	A	0.67	0.24	0.06	0.02
	B	0.52	0.33	0.11	0.05
$35-54$	A	0.70	0.22	0.05	0.02
	B	0.54	0.31	0.10	0.04
55+	A	0.84	0.12	0.02	0.01
	B	0.68	0.24	0.06	0.02
$\Delta P(Y=1)$					
$18-34$		-0.15***	0.08***	0.05***	0.03***
		(0.04)	(0.02)	(0.01)	(0.01)
$35-54$		-0.16***	0.09***	0.05***	0.02***
		(0.04)	(0.02)	(0.01)	(0.01)
55+		-0.17***	0.12***	0.04***	0.02***
		(0.03)	(0.02)	(0.01)	(0.00)

Table A-12: Interaction ordered logistic regression: Age x treatment, standard covariates (full table)

	(1) Number correct	(2) Number incorrect	(3) Number DK
Group B	-0.44***	-0.10	0.75***
	(0.17)	(0.19)	(0.19)
35-54	0.78***	-0.88***	-0.16
	(0.16)	(0.19)	(0.19)
55+	1.67***	-1.41***	-1.11***
	(0.18)	(0.19)	(0.20)
Group B × 35-54	-0.35	0.46*	0.03
	(0.24)	(0.27)	(0.27)
Group B × 55+	-0.27	0.06	0.33
	(0.23)	(0.28)	(0.27)
Covariates	All	All	All
/cut1	-2.23***	-0.00	0.47**
	(0.21)	(0.22)	(0.22)
/cut2	-0.46**	2.01***	2.30***
	(0.20)	(0.23)	(0.23)
/cut3	1.05***	3.99***	3.65***
	(0.20)	(0.31)	(0.28)
Observations	3502	3502	3502

Note: The Stata command `ologit` is used to generate these tables. "/cut" corresponds to cutoff points of a latent continuous variable, $Num*$, which determine the observed count variable (for example, $NumDK$).

the direction of the effect is positive, it is not statistically significant for the majority of college-educated respondents when compared with high school students.

In contrast, Figure A-6 shows the interaction effects for university-educated respondents relative to high school ones. Here, the interaction effect is negative for university graduates when compared with high school graduates and is also statistical significant (z-scores are mostly below the threshold of -1.96. This suggests that, relative to high school graduates, university graduates have significantly lower probability of answering DK when moving from Group A to Group B.

Table A-13: Results from logistic regression, education x treatment

			Correct (> 0)	Incorrect (> 0)	Don't know (> 0)
$P(Y = 1)$	High school	A	0.95	0.36	0.33
		B	0.88	0.36	0.52
	College	A	0.98	0.31	0.22
		B	0.92	0.30	0.40
	University	A	0.98	0.16	0.16
		B	0.97	0.22	0.24
$\Delta P(Y = 1)$	High school		-0.07*** (0.02)	0.01 (0.04)	0.19*** (0.04)
	College		-0.06*** (0.02)	-0.01 (0.03)	0.19*** (0.03)
	University		-0.01* (0.01)	0.06** (0.02)	0.09*** (0.03)

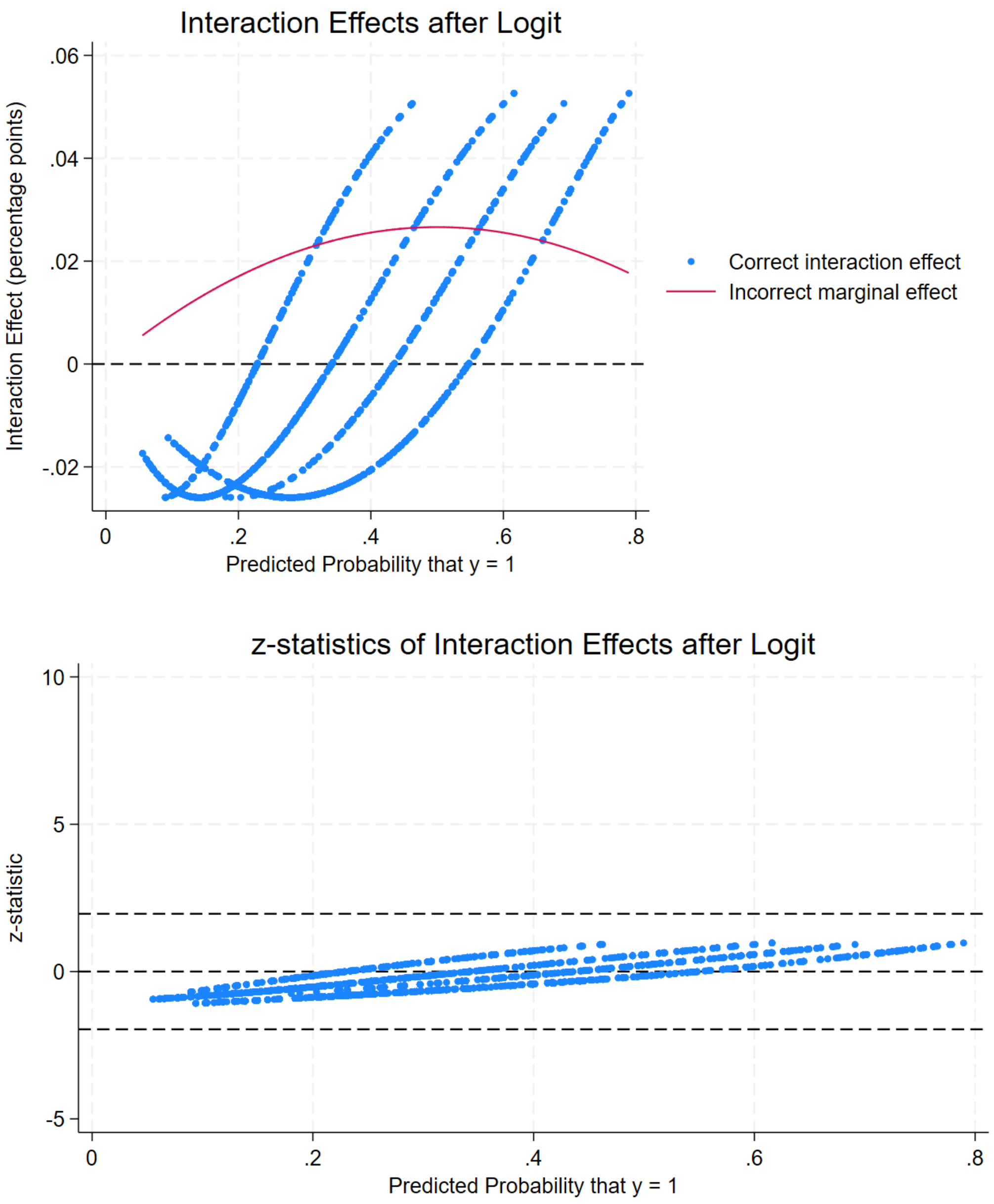
Note: *** p <0.01, ** p<0.05, * p < 0.1. $P(Y = 1)$ refers to the estimated probability of each group having outcome Y (> 0 correct, incorrect, or DK responses). $\Delta P(Y = 1)$ corresponds to the treatment effect, estimated separately by highest education achieved. Logistic regression specification includes an interaction term between education (group indicator) and treatment.

A.6. Heterogeneity analysis: Uncertainty

While we do not include it as a core covariate, we also perform heterogeneity analysis using an additional grouping: respondents who demonstrate uncertainty in other survey questions, and respondents who do not. This gives us a proxy for the propensity to opt out of survey questions by answering "Don't know," allowing us to further investigate which type of respondent may be most affected by fatigue.

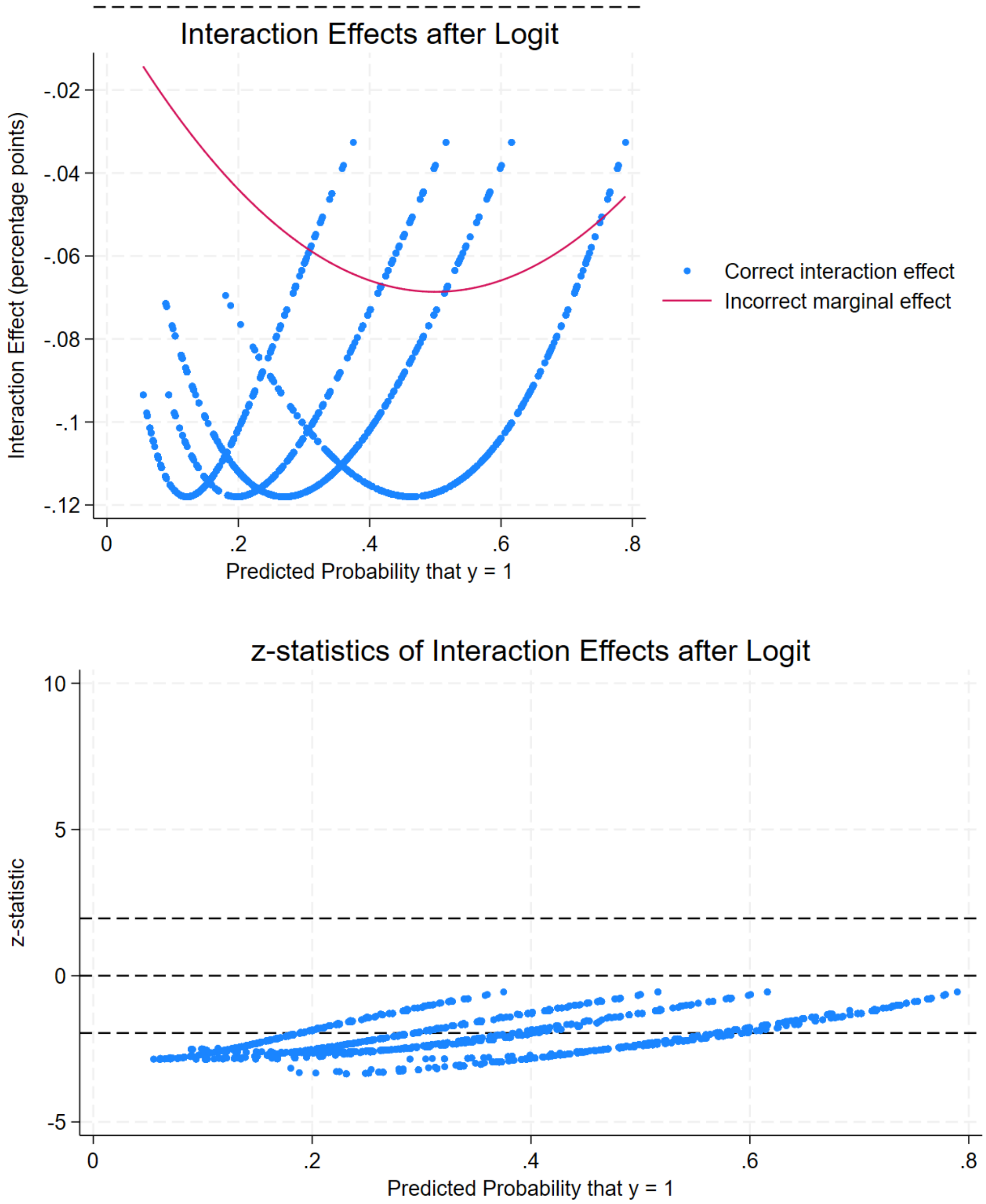
We define a new variable, *Uncertain*, which takes the value of 1 if a respondent answers "Don't know" or "Not sure" to any other survey question, and 0 otherwise. There are three other (non-FL) survey questions that have "Don't know" or "Not sure" as options:

Figure A-5: Interaction effects: Education (college), any "Don't know" (DK) responses



Note: Education (high school) is used as the base category in this case. These plots show the estimated interaction effects by probability of responding DK to any of the three questions, as specified by Ai et al. (2004). The command *inteff* is used.

Figure A-6: Interaction effects: Education (university), any "Don't know" (DK) responses



Note: Education (high school) is used as the base category in this case. These plots show the estimated interaction effects by probability of responding DK to any of the three questions, as specified by Ai et al. (2004). The command *inteff* is used.

Table A-14: Results from ordered logistic regression, education x treatment

		Number correct				Number incorrect			
		0	1	2	3	0	1	2	3
$P(Y=1)$									
High school	A	0.06	0.19	0.31	0.44	0.64	0.28	0.07	0.01
	B	0.11	0.27	0.31	0.30	0.63	0.29	0.07	0.01
College	A	0.04	0.14	0.27	0.56	0.69	0.25	0.05	0.01
	B	0.07	0.20	0.31	0.42	0.70	0.24	0.05	0.01
University	A	0.02	0.07	0.19	0.71	0.83	0.14	0.02	0.00
	B	0.03	0.12	0.25	0.60	0.78	0.18	0.03	0.01
$\Delta P(Y=1)$									
High school		0.05***	0.08***	0.01	-0.14***	-0.01	0.00	0.00	0.00
		(0.01)	(0.02)	(0.01)	(0.03)	(0.04)	(0.03)	(0.01)	(0.00)
College		0.03***	0.07***	0.04***	-0.14***	0.01	-0.01	0.00	0.00
		(0.01)	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.01)	(0.00)
University		0.01***	0.04***	0.06***	-0.11***	-0.05**	0.04**	0.01**	0.00*
		(0.00)	(0.01)	(0.01)	(0.03)	(0.02)	(0.02)	(0.00)	(0.00)

		Number DK			
		0	1	2	3
$P(Y=1)$					
High school	A	0.67	0.25	0.06	0.02
	B	0.48	0.35	0.12	0.05
College	A	0.78	0.17	0.03	0.01
	B	0.60	0.29	0.08	0.03
University	A	0.84	0.13	0.02	0.01
	B	0.75	0.19	0.04	0.01
$\Delta P(Y=1)$					
High school		-0.19***	0.10***	0.06***	0.03***
		(0.04)	(0.02)	(0.01)	(0.01)
College		-0.19***	0.12***	0.05***	0.02***
		(0.03)	(0.02)	(0.01)	(0.00)
University		-0.09***	0.07***	0.02***	0.01***
		(0.03)	(0.02)	(0.01)	(0.00)

Table A-15: Interaction ordered logistic regression: Education x treatment, standard covariates (full table)

	(1) Number correct	(2) Number incorrect	(3) Number DK
Group B	-0.69***	0.04	0.89***
	(0.16)	(0.19)	(0.18)
College	0.56***	-0.26	-0.64***
	(0.16)	(0.18)	(0.18)
University	1.33***	-1.10***	-1.06***
	(0.17)	(0.20)	(0.20)
Group B × College	0.05	-0.09	0.10
	(0.22)	(0.25)	(0.24)
Group B × University	0.12	0.32	-0.28
	(0.22)	(0.25)	(0.25)
Covariates	All	All	All
/cut1	-2.35***	0.07	0.54**
	(0.21)	(0.23)	(0.22)
/cut2	-0.58***	2.08***	2.37***
	(0.20)	(0.24)	(0.23)
/cut3	0.93***	4.06***	3.72***
	(0.21)	(0.31)	(0.27)
Observations	3502	3502	3502

Note: The Stata command `ologit` is used to generate these tables. "/cut" corresponds to cutoff points of a latent continuous variable, $Num*$, which determine the observed count variable (for example, $NumDK$).

- There have been reports of a perceived risk of viruses being transmitted through the handling of cash. Compared with before the COVID-19 pandemic began, how has your use of cash changed, if at all, in light of these reports?

- In the past week, have you heard of, seen or experienced a store or business refusing to accept cash due to the perceived risk of viruses being transmitted?

- Would it be problematic for you if Canadian consumers stopped transacting with cash or businesses stopped accepting cash?

Table A-17 shows the interaction regressions with treatment and uncertainty. Respondents who are uncertain provide fewer correct FL responses (0.21 fewer; $p < 0.01$), and more DK responses (0.19 more; $p < 0.01$). Uncertain respondents also have a larger treatment effect on their number of DK answers, and a smaller treatment effect on the number of incorrect responses—but no difference in the effect on the number of correct responses. Altogether, there is evidence that these individuals are more affected by fatigue.

Table A-16: Interaction ordered logistic regression: Uncertain x treatment, standard covariates (full table)

	(1) Number correct	(2) Number incorrect	(3) Number DK
Group B	-0.62***	0.19	0.81***
	(0.11)	(0.12)	(0.13)
Uncertain	-0.51***	0.05	0.71***
	(0.17)	(0.19)	(0.18)
Group B × Uncertain	-0.09	-0.57**	0.32
	(0.23)	(0.29)	(0.25)
Covariates	All	All	All
/cut1	-2.43***	0.09	0.64***
	(0.21)	(0.22)	(0.21)
/cut2	-0.64***	2.11***	2.52***
	(0.20)	(0.23)	(0.22)
/cut3	0.88***	4.09***	3.90***
	(0.20)	(0.31)	(0.26)
Observations	3502	3502	3502

Note: The Stata command `ologit` is used to generate these tables. "/cut" corresponds to cutoff points of a latent continuous variable, Num*, which determine the observed count variable (for example, $NumDK$).

Table A-17: Regression: Number of correct, incorrect, and "Don't know" responses on treatment and "uncertain" group

	Correct		Incorrect		Don't know	
	#	> 0	#	> 0	#	> 0
Treatment	-0.25***	-0.05***	0.04	0.04	0.21***	0.14***
	(0.04)	(0.01)	(0.03)	(0.02)	(0.03)	(0.02)
Uncertain	-0.21***	-0.04***	0.02	0.01	0.19***	0.12***
	(0.07)	(0.02)	(0.05)	(0.03)	(0.05)	(0.04)
Treatment × Uncertain	-0.08	-0.01	-0.14**	-0.10**	0.22**	0.08
	(0.10)	(0.03)	(0.07)	(0.05)	(0.09)	(0.05)
Controls?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,496	3,496	3,496	3,496	3,496	3,496

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Number of correct responses is defined as the total number of correct responses. Number of incorrect and "Don't know" is defined the same way. Treatment takes on a value of 1 if an individual is in Group B. *Uncertain* is an indicator taking the value of 1 if the respondent answered "Don't know" or "Unsure" to any other survey questions in the CAS (of which there are three). Covariates include the standard set: age, gender, region, education, income, marital status, and having kids at home.

A.7. Oaxaca–Blinder using each STC residual

The CAS sample was determined through quota sampling, with nested sampling targets by age, gender, and region (Chen et al., 2021). While the survey groups A and B are randomly assigned, it is still possible that the two sub-samples differ along unobservable characteristics (for example, if survey attrition rates are higher among respondents who begin with the FL questions). Another potential concern is the question of whether the CAS sample, more broadly speaking, is truly representative of the population, given the non-random sampling procedure.

One robustness check that we use to address these concerns is comparing our CAS subsamples with external data. The CAS includes thirteen questions borrowed from the Canadian Perspectives Survey Series (CPSS) 4 and 5, Statistics Canada surveys conducted using probability sampling.[3] These questions are related to COVID-19 behaviors (masking, preparing for the pandemic, and more) as well as internet security precautions (not letting websites remember personal information, shopping only on reputable websites, using strong passwords, and more). The questions are all Yes/No, and are coded as binary variables. We take the overlapping questions from the CAS and CPSS and perform additional analysis using them.

Using the Statistics Canada micro data, we first construct weighted means by age and gender

[3] The samples for the CPSS are composed of Labour Force Survey respondents, and participation in the CPSS is voluntary.

group (18-34, 35-54, 55+; male, female), denoted as $\bar{S}^{STC}_{A,G}$. Using the CAS data, we then subtract these mean values for each overlapping question, to get a measure of deviation from probability-sampled expected values:

$$\tilde{S}_i = S^{CAS}_i - \bar{S}^{STC}_{A,G}.$$

We then regress $\tilde{S}_i$ on the full set of CAS covariates, but not treatment (age, gender, region, income, education, marital status, having kids at home):

$$\tilde{S}_i = \beta X_i + \epsilon_i.$$

The goal of this regression is to predict the error term, $\hat{\epsilon}_i$, extracting the portion of detrended responses which is orthogonal to the standard CAS covariates. If CAS respondents differ largely from Statistics Canada CPSS respondents on unobservable differences, then we can use these de-trended error measures as a proxy for the divergence. Furthermore and more importantly, we can assess whether the divergence differs by A/B group.

Table A-18 shows the results when adding these predicted error measures to the standard pooled Oaxaca–Blinder decompositions. From this analysis, it appears that adding the predicted error terms only slightly increases the explained share, and does not change the insignificance. From this, we might conclude that much of the A/B treatment effect is not explained by differences in deviations from the Statistics Canada means.

Table A-18: Oaxaca–Blinder decomposition: With residuals

			Correct			*Don't know*	
	FL score	Interest	Inflation	Risk	Interest	Inflation	Risk
Difference	-0.32***	-0.08***	-0.06**	-0.16***	0.05***	0.07***	0.15***
Explained	-0.06**	-0.01*	-0.02**	-0.02*	0.00*	0.01*	0.01*
Unexplained	-0.25***	-0.07***	-0.04*	-0.14***	0.04***	0.06***	0.14***
Explained ϵ	Cleanliness	None	None	None	None	None	Didn't shop online
Unexplained ϵ	None						

A.8. Measuring Financial Literacy

The main results of our analysis suggest that measuring FL is not as simple as benchmarking on the number of correct responses given by a survey respondent, as the placement of FL questions can dictate the types of responses given. With this problem in mind, we investigate several alternative methods to measure financial literacy.

Figure A-7 provides a demonstration of the performance of several common indices. The first (FL score) is an index defined using the following formula, which does not directly account for DK responses: $FLScore = N(correct) - N(incorrect)$. In this formula, each correct response is awarded one point, and each incorrect response results in a point subtracted. DK responses are assigned a value of zero, which is equivalent to considering their "true" response to be a fair coin toss between correct and incorrect.[4] In reality, more than half of respondents who answer DK may in fact know the correct response. Bucher-Koenen et al. (2025) show that between 60% and 70% of Dutch respondents who answer "Don't know" can correctly answer the Big Three FL questions later, when the DK option is removed. In practice, using a score like this that penalizes DK responses could artificially lower measured FL in the case of survey fatigue. Using this measurement of financial literacy, we find that Group B has an average FL score of about 16% lower than Group A. Considering only the average number of correct responses is similarly flawed. Another measure used to measure financial literacy is reporting the percentage of respondents who answer all three questions correctly. Using this measure, Group A still outperforms Group B.

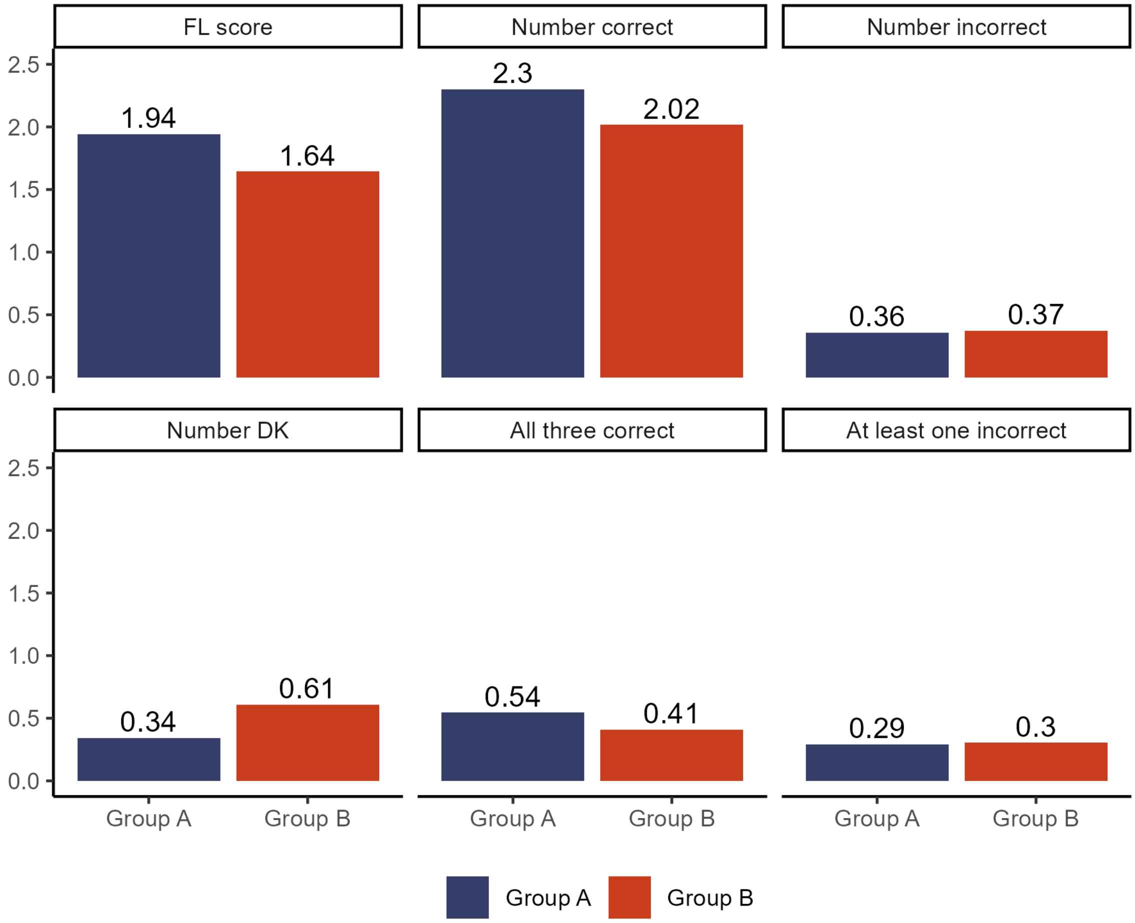
As survey fatigue tends to result in a higher proportion of DK responses and a lower proportion of correct responses, another alternative is to report on the number of incorrect responses. **Figure A-7** shows that this may be the most robust option to measure financial literacy, as the average number of incorrect responses is similar in groups A and B (0.36 for A, and 0.37 for B). The percentage of respondents who got at least one question incorrect is also similar between the groups (29% in A vs. 30% in B).

To showcase the potential effect this can have on reported FL, we consider the Methods-of-Payment (MOP) Survey, a survey run annually by the Bank of Canada since 2021. The MOP is similar in design to the CAS. Between 2023 and 2024, the MOP underwent a change in survey design, where the three FL questions were moved from the beginning of the survey to midway through. **Figure A-8** illustrates the associated decline in performance. Between 2023 and 2024, the average number of correct responses fell rather sharply (-0.26), while the number of DK responses rose (+0.18). The number of incorrect responses did also rise slightly, though not as starkly (+0.08). The percentage of respondents answering all three questions correctly fell by 11 percentage points, whereas the percentage of respondents answering at least one question incorrectly rose by 3 percentage points. This analysis does not consider that alternative time trends could be driving the change between 2023-24, but points to the possibility that the ordering of FL questions could influence how we perceive FL over time, with different indices/benchmarks carrying varying levels of robustness.

Ultimately, the most comprehensive way to measure FL is to consider the full picture: To measure and report on correct, incorrect, and DK answers separately, and to be transparent about the survey design.

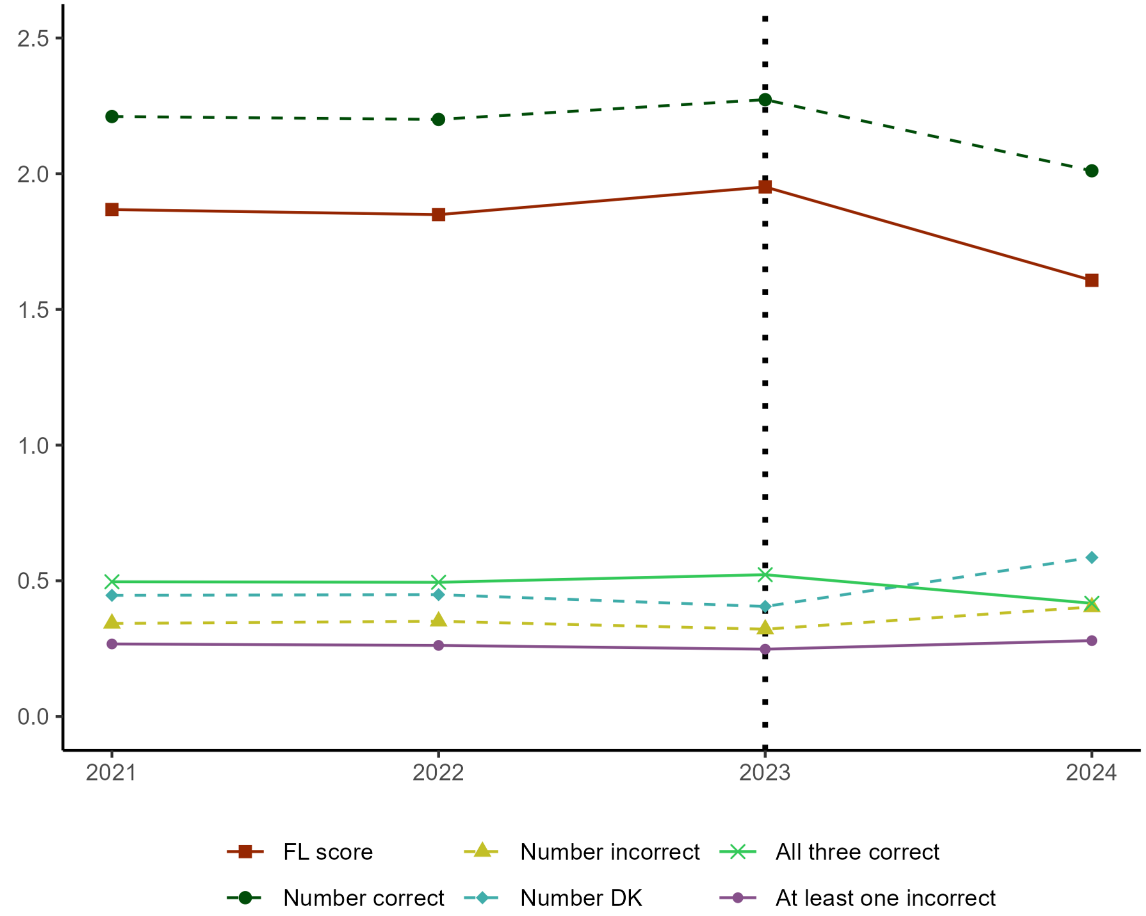
[4] Consider the expected value of an individual's true response, given that they select DK: $E[correct|DK = 1] = (1)P(correct|DK = 1) + (-1)P(incorrect|DK = 1) = 2P(correct|DK = 1) - 1$. Setting $E[correct|DK = 1] = 0$ implies that $P(correct|DK = 1) = \frac{1}{2}$.

Figure A-7: Comparison of FL reporting indices: Group A vs. Group B



Note: This figure compares the means of various measures of financial literacy, compared across Group A (questions at the beginning of the survey) and Group B (questions at the end). *FL score* refers to a measure defined using the number of correct, incorrect, and DK responses: $N(correct) - N(incorrect) + 0 * N(DK)$. *Number correct* refers to the overall number of correct FL responses. *Number incorrect* and *Number DK* similarly report differences in the mean number of incorrect and DK responses. *All three correct* is a measure taking the value of 1 if a respondent answers all three FL questions correctly, and *At least one incorrect* takes the value of 1 if a respondent answers at least one FL question incorrectly.

Figure A-8: Comparison of FL reporting indices: Using the Methods-of-Payment (MOP) Survey



Note: This figure compares the means of various measures of financial literacy, compared over time. In the 2024 MOP Survey, the FL questions were moved from the beginning to midway through the survey. *FL score* refers to a measure defined using the number of correct, incorrect, and DK responses: $N(correct) - N(incorrect) + 0 * N(DK)$. *Number correct* refers to the overall number of correct FL responses. *Number incorrect* and *Number DK* similarly report differences in the mean number of incorrect and DK responses. *All three correct* is a measure taking the value of 1 if a respondent answers all three FL questions correctly, and *At least one incorrect* takes the value of 1 if a respondent answers at least one FL question incorrectly.