Australian Government
Productivity Commission

# About the survey and the results

Superannuation:
Alternative Default Models
*Supplement to the Draft Report*

July 2017

### An appropriate reference for this publication is:

Productivity Commission 2017, 'About the survey and the results', Supplement to the Draft Report, *Superannuation: Alternative Default Models*, Canberra.

### Publications enquiries

Media and Publications, phone: (03) 96 53 2244 or email: maps@pc.gov.au

### The Productivity Commission

The Productivity Commission is the Australian Government's independent research and advisory body on a range of economic, social and environmental issues affecting the welfare of Australians. Its role, expressed most simply, is to help governments make better policies, in the long term interest of the Australian community.

The Commission's independence is underpinned by an Act of Parliament. Its processes and outputs are open to public scrutiny and are driven by concern for the wellbeing of the community as a whole.

Further information on the Productivity Commission can be obtained from the Commission's website (www.pc.gov.au).

# Contents

# About the survey and the results

### Key points

- As part of its inquiry into Alternative Default Allocation Models, the Commission undertook a survey of member behaviour. Respondents were surveyed on their past and present experiences and attitudes towards super, and also asked to complete some experimental tasks to better understand how they make their decisions.
- Members' responses about their existing fund suggest that demographic characteristics, such as age and household income, generally seem to have little influence on a person's propensity to default. However, individuals from households with high incomes are less likely to default than those with lower incomes, and males are slightly less likely to default than females.
- When asked how they would go about selecting a new fund, respondents tended to rank comparison websites and key performance indicators as the most important sources of information. Employers, and friends and family were ranked the least important. This result is stronger for older respondents and those that scored highly on the financial literacy tests.
- Respondents tended to rank net investment returns and fees as more important product features than member services. This result is stronger for older respondents and those that scored highly on the financial literacy tests.
- The survey also contained a choice experiment in which respondents were asked to nominate a super fund. One group was assisted with an optional shortlist of funds with their key metrics (assisted choice), while another was provided with simply an empty text box (unassisted choice).
- For both groups, the probability of failing to nominate any fund was low — under 20 per cent for unassisted choice, and under 5 per cent for the group that was assisted.
  - However, in the unassisted choice group, respondents had a tendency to rely on selecting their existing fund, whereas respondents in the assisted choice group fund were more likely to select a new fund from the shortlist.
  - Reducing the number of options on the shortlist from eight to four appeared to reduce the difficulty of the task reported by respondents, but the effect was small. A low level of financial literacy is a much stronger predictor of increased perceived difficulty.
- Respondents sought to maximise investment returns, minimise fees, and minimise risk in their nomination decisions when they chose from the shortlist. Despite this, there appears to be substantial heterogeneity in choices, which might be driven by other factors, for example recognition of a fund's brand.

To assist with the development and assessment of alternative default models, the Commission undertook an experimental survey of the general public (PC 2017, chapter 4). This supplement provides information about the survey and presents results. It is structured as follows: section 1 outlines the survey's design and implementation; section 2 explores the data — in particular, the sample profile and general usability concerns; section 3 presents results on how respondents chose their current superannuation product; section 4 presents results on respondents' attitudes towards superannuation; and sections 5 to section 7 present results from the choice experiment (the experiment itself is explained in section 1).

The results detailed in this supplement will be used to assist in refining the models. Explaining these results inevitably involves the use of statistical and other jargon. Box 1 below contains some of the key definitions.[1]

Box 1 **Key definitions**

*Randomised control trial (RCT)* — experimental methods used to evaluate the effects of interventions on a particular outcome. *Randomised* refers to the fact that subjects are randomly allocated between a *control group* (which do not receive the intervention) and a *treatment* group (which receive the intervention).

*Balance*d *distribution* — distribution of subjects (on various demographics) in the control group of a RCT is similar to the distribution of subjects in the treatment group.

*Economic significance* — the magnitude of the estimated value. In this context, where an estimated effect has a material size, it is considered to be economically significant.

*Power* — the sample sizes for a particular analysis are large enough to identify a treatment effect, if there is one.

*Regression Analysis* — a broad group of statistical techniques which seek to explain an outcome of interest as a function of variables which are observed by the analyst, such as demographics. *Probit* regressions are employed when the outcome of interest is an *indicator variable.* These are variables (for example, whether a respondent has nominated a fund) which take on values of zero or one, with each representing a particular case (one if someone has nominated, zero otherwise). Sets of indicators can also be used to represent discrete variables, for example the household income bracket of a respondent.

*Standard error* — the variability of the estimate. Larger standard errors place greater uncertainty on the estimates.

*Statistical significance* — in this analysis, an estimate is said to be statistically significant if the p-value is less than 10 per cent.

*Treatment effects* — based on a RCT, the (average) treatment effect for an intervention on an outcome can be estimated for the population. The treatment effect is computed as the average outcome in the treatment group minus the average outcome in the control group. Additional insight can be gained by considering *varying treatment effects*. That is, how does the effect of an intervention change depending on the demographic of the subjects? The difference between an average treatment effect for the entire pool and the average treatment effect for a particular group of people (for example, a more financially literate group) is often referred to as an *interaction effect*.

1 The Commission is grateful to Dr. Andrew Reeson of CSIRO's Data61 for his helpful comments on a draft of this supplement.

# 1 About the survey

The survey was primarily designed to gather evidence relevant to understanding how people would likely behave under the baseline scenario (unassisted choice), as well as in the presence of a shortlist of superannuation products (assisted choice). The survey was also intended to address gaps in the evidence already available from other surveys and behavioural research. The survey was not intended to be used for evaluating current policy settings.

Given the specialised nature of the research, the Commission sought external expertise, particularly in behavioural finance and Randomised control trials (RCTs) to design and conduct the survey. Insight Analytics was engaged to design the survey on the Commission's behalf. Insight Analytics also conducted the survey, in conjunction with a third-party panel provider. Analysis was conducted internally by the Commission. The survey may be used for future research by the Commission.

The survey was conducted online in early 2017, with a target quota of 2000 complete responses. The survey questionnaire, data and associated documentation will be made available for download at a later date on the Commission's website at http://www.pc.gov.au/inquiries/current/superannuation/alternative-default-models/draft.

## Survey design

To gather relevant data for the inquiry, the Commission developed several research questions:

- how do people currently choose their superannuation fund and product? (sections 3 and 4)
- how many people would nominate a product under the Commission's baseline of unassisted choice, and what factors would influence their decision? (section 5)
- how do choices change when people are given a non-binding shortlist of superannuation products — and relevant information on each product — to choose from? (section 6 and 7)

The survey contained two types of questions: general questions to gather information on demographic characteristics of respondents, their past experiences and financial literacy; and experimental questions (figure 1). The experimental questions were designed to elicit information about respondents' decision making and behaviour by assigning different respondents to different 'treatments'. Treatment design varied across groups of respondents to test the impact of specific presentation elements.

The survey was estimated to have taken most respondents about 10 to 15 minutes to complete. The main components are discussed below, in the order in which they appeared.

Figure 1 **Components of the survey**

**General questions**
non-experimental

- Screening
- Current state
- Closed-ended
- Demographics

**Choice experiment**

- Questions about future behaviour (ex ante)
- Hypothetical
- Control and treatment groups

**List experiment**

- Questions about past behaviour (ex post)
- Real life
- Control and treatment groups

### Screening questions

Survey respondents were first asked several screening questions to determine their eligibility to complete the survey. These questions covered gender, age group and place of residence. Pre-set target quotas were imposed (based on Census data) to ensure a representative sample along these characteristics. Once each quota was exceeded, no more participants with those characteristics were able to participate in the survey.

### Current state questions

Participating respondents were then asked several questions about their current superannuation arrangements. Specifically, they were asked whether they had (or had ever had) an account with a superannuation fund, the name of their current fund, and how long they had been with that fund. Throughout the survey, no distinction was drawn between funds and products, and 'fund' was used in place of 'product' to minimise confusion for respondents.

Participants were also asked how they selected their current fund. This was done in open-response format to minimise any biases that could arise from framing or social desirability effects (which can arise where participants respond based on a list of predefined options). The placement of this question early in the survey, before the experimental components, was also designed to minimise any potential biases arising from the survey questions themselves.

## Choice experiment

For this experiment, respondents were randomly assigned to one of two main treatments: 17 per cent to unassisted active choice (the 'control group'), and the remaining 83 per cent to assisted employee choice (the 'treatment group'). Respondents were not told that they had been randomly assigned to different treatments. More respondents were assigned to the treatment group due to the number of variations being tested within that group (described below). Respondents were requested to imagine they were starting a new job and had to choose their own superannuation fund; they were asked 'what super fund would you go with?'.

Respondents in the control group faced a nomination decision without any assistance. Respondents in the treatment group were provided a shortlist of funds to assist in their nomination. The shortlist included four metrics for each fund, covering the risk level,[2] past returns, return target and fees associated with it. Within the group, respondents were assigned to one of ten specific treatment categories (figure 2). Specifically, each respondent was presented with either four, five, six, seven or eight options, and with the fee and return metrics presented either in terms of the percentage of account balance or in dollar figures (based on a nominal $50 000 balance). Respondents also had the option to select 'something else' and nominate a different fund in a free-text entry box.

The hypothetical products used for this exercise were loosely based on a selection of real MySuper products available in the market. In each instance, respondents were presented with a set of actual fund logos combined with a block of the four metrics (returns, fees, risk, and target). The eight funds used in the experiment were chosen based on their actual size (total membership and assets under management), with the blocks of metrics based on real performance figures for these funds' MySuper products. The set comprised both industry and retail funds.

Within each of the ten treatment categories, there was further random assignment of the:

- order in which funds were listed
- specific fund name and corresponding logo
- block of metrics shown for each fund.

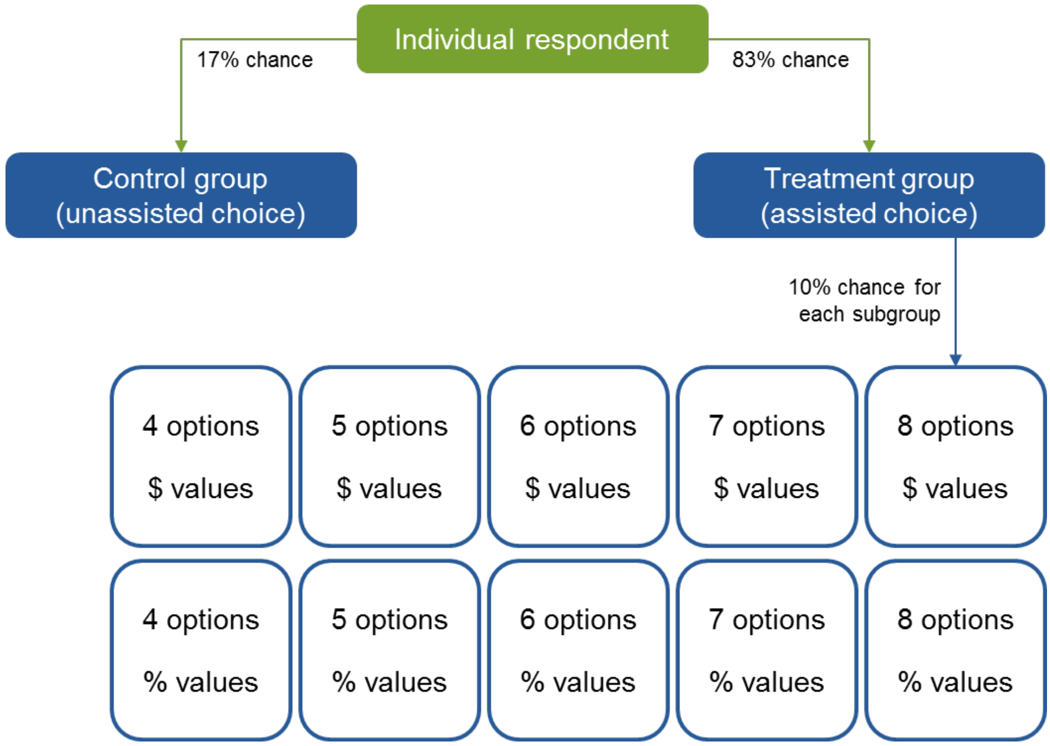
These elements were randomised to allow for subsequent testing of (or control for) the effects of ordering, presentation and brand sentiment on decision making.

In addition to nominating a product on the shortlist, respondents were also able to nominate a fund of their choosing or not nominate at all. Once respondents had completed the experiment, they were asked why they chose the particular fund, in a free text field.

2 Risk was presented using the Australian Prudential Regulation Authority standard risk measure. It was chosen due to its consistent application across MySuper product dashboards and Product disclosure statements.

They were also asked to explain how difficult they found the exercise, and to score this on a scale of one to five (where five was the most difficult).

Figure 2 **Choice experiment: control and treatment groups**



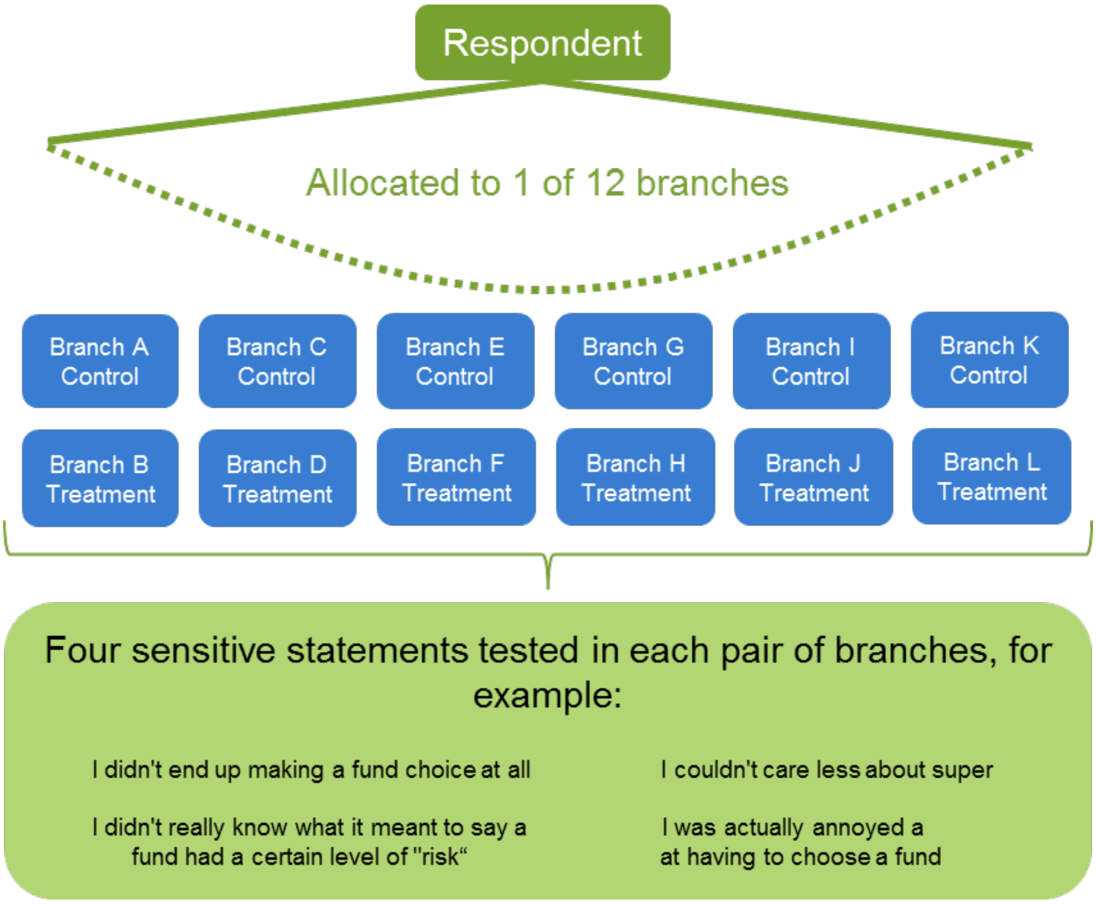

## List experiment

A 'list experiment' is a technique that is commonly used by psychologists and political scientists to gauge attitudes where respondents may not feel comfortable providing honest answers to questions asked directly. To overcome potential biases that can arise from social desirability, non-response and other effects, the questions of interest are asked indirectly by combining them with a set of other, less sensitive questions.

This list experiment was designed to elicit information on behaviours, decision making and attitudes related to the last time that respondents had to choose a superannuation fund (those who had never had a fund were excluded from the task) (figure 3).[3] Respondents were randomly assigned to one of 12 'branches'. In each branch, four lists were shown with the treatment branches containing the sensitive statement. The order of statements in each list was randomised for each respondent. Full details about the list experiment can be found in annex A.2.

3 The full list of sensitive statements tested can be found in tables 2 and 3.

Figure 3 **List experiment design**




### Closed-ended and demographic questions

The remainder of the survey contained several closed-ended questions. First, respondents were asked a series of three multiple-choice financial literacy questions (based on those used by Lusardi and Mitchell (2011) and OECD (2011)), and an additional question testing their understanding of the Australian superannuation system.

Next, respondents were asked to complete two ranking exercises. In the first, they had to rank from most to least important four specific factors when selecting a superannuation fund (fees, net returns, member services and choice of investment options). In the second, they had to rank sources of information relevant to selecting a superannuation fund from most to least helpful (key performance indicators, product disclosure statements, comparison websites, financial advisers, friends or family, and employers).

Following this, respondents were asked about what they did the last time they had to choose a superannuation fund, and could select from several preset options or could enter their own free text (the question was identical to an open-response question earlier in the survey).

The survey finished with several demographic questions to gather more information on respondents' characteristics. These covered postcode, labour force status, occupation, education level and household income.

## Survey conduct

Insight Analytics conducted the survey online from 22 February to 2 March 2017, via a third-party panel provider (Quality Online Research). Respondents were rewarded for their participation with $1 per every five minutes spent on the survey, plus entry into a biannual prize draw. The final survey data and documentation were delivered to the Commission on 16 March 2017, following some cleaning and checking of the data by Insight Analytics.

The full sample included 2351 respondents, with approximately 2000 completing all major components of the survey.[4] Most analyses use a smaller subset of responses, where data quality problems preclude use of the full set and to approximate ideal experimental conditions (section 2 and annex A.5).The survey was successful in achieving the Census quotas, resulting in a full sample broadly representative of the working-age Australian population (figures 4 and 5).[5]

## Interpreting the survey results

Different methods of analysis have been used to analyse the survey data, depending on the suitability of each method for addressing the relevant research question(s). The results are not intended to be used for evaluating current policy settings. Most of the analysis is exploratory: many results are descriptive without causal interpretations, since correlation is not necessarily causation. That is, the results may provide potential insights, as opposed to facts, about how superannuation decisions might be made in real life.

This cautious approach is warranted. Causal inference is difficult and requires carefully designed and executed RCTs to be fully valid.[6] In the experimental treatments, without having a good idea of the underlying mechanisms, causal interpretations can still be problematic.

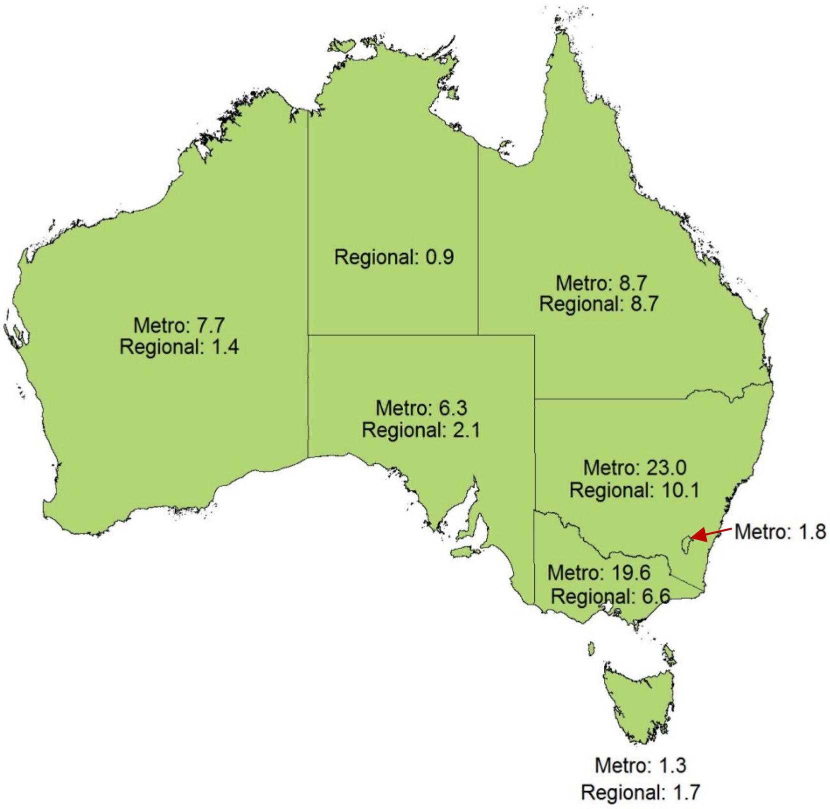
4 Three observations are exact duplicates, and have been removed from the data. This means, the full sample is really 2348 respondents.

5 Unless otherwise stated, all data are sourced from the Commission's survey.

6 See Rubin (2008) for a discussion about the concept of approximating experimental methods and Gordon et al. (2017) showing the importance of using experimental methods.

Figure 4 **Sample profile by region**[a]

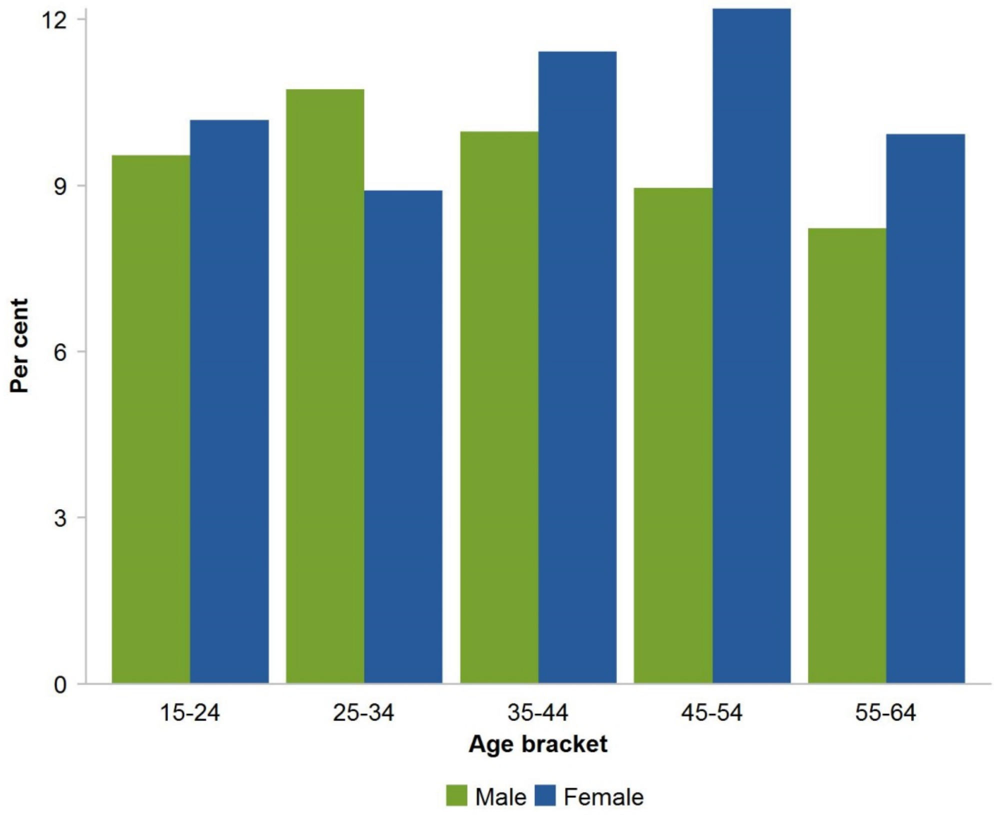
N = 2347




**a** One observation lacked the associated region. **b** Figures are reported as percentages, which may not sum to 100 due to rounding.

Figure 5 **Sample profile by gender and age[a]**

N = 2348




## 2 Suitability of data and transformations

Prior to commencing analysis the Commission prepared the data by cleaning, rearranging and creating new variables of interest. Suitability of the data for analysis was also examined. This section commences by illustrating the representativeness of the sample profile on a variety of demographics and observables — some of which have been constructed. Some other features of the data are highlighted. The section concludes with a discussion of the suitability of the data for analysis.

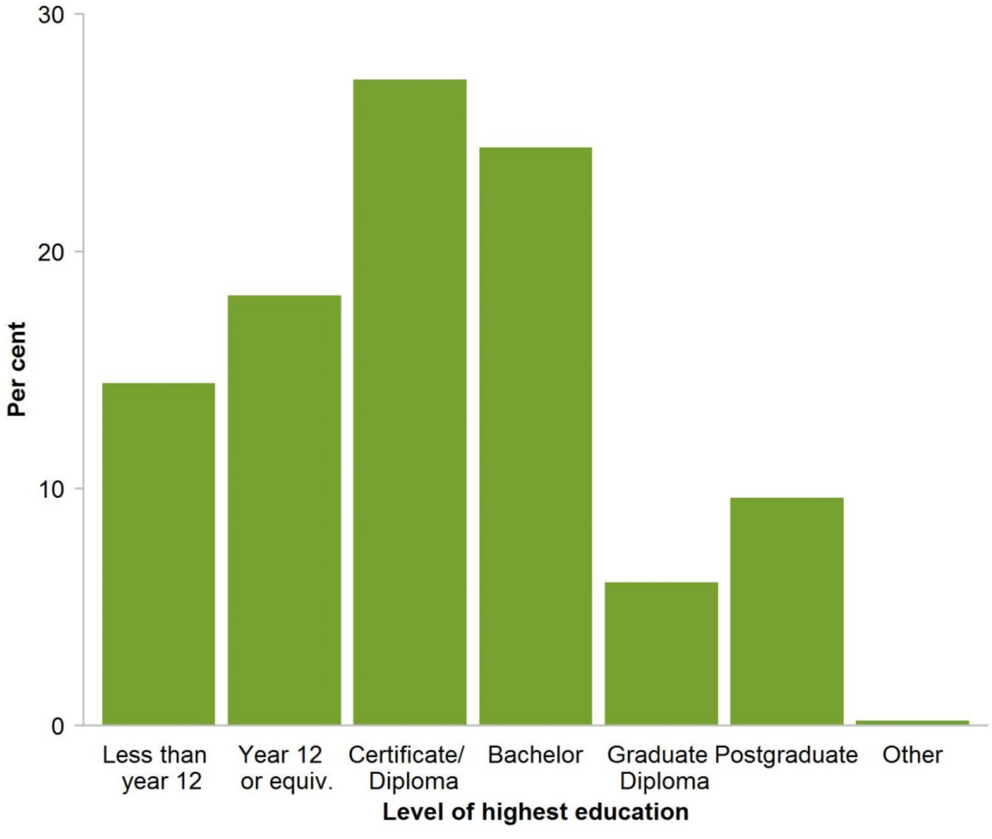
## Sample profile and additional observables

### Level of highest education

Respondents were asked about the highest level of education they had completed (figure 6). The sample is broadly reflective of the education status of the wider working-age Australian population (ABS 2016).

Figure 6 **Sample profile by level of highest education[a,b]**

N = 2002




[a] 'Less than year 12' is the categories 'primary school', 'some high school' and 'year 10 or equivalent' combined. 'Certificate/Diploma' is 'Trade certificate/Apprenticeship' and 'Diploma/Associate Degree' combined. 'Postgraduate' is 'Masters' and 'PhD' combined. [b] There were 346 non-responses as this question was not mandatory.

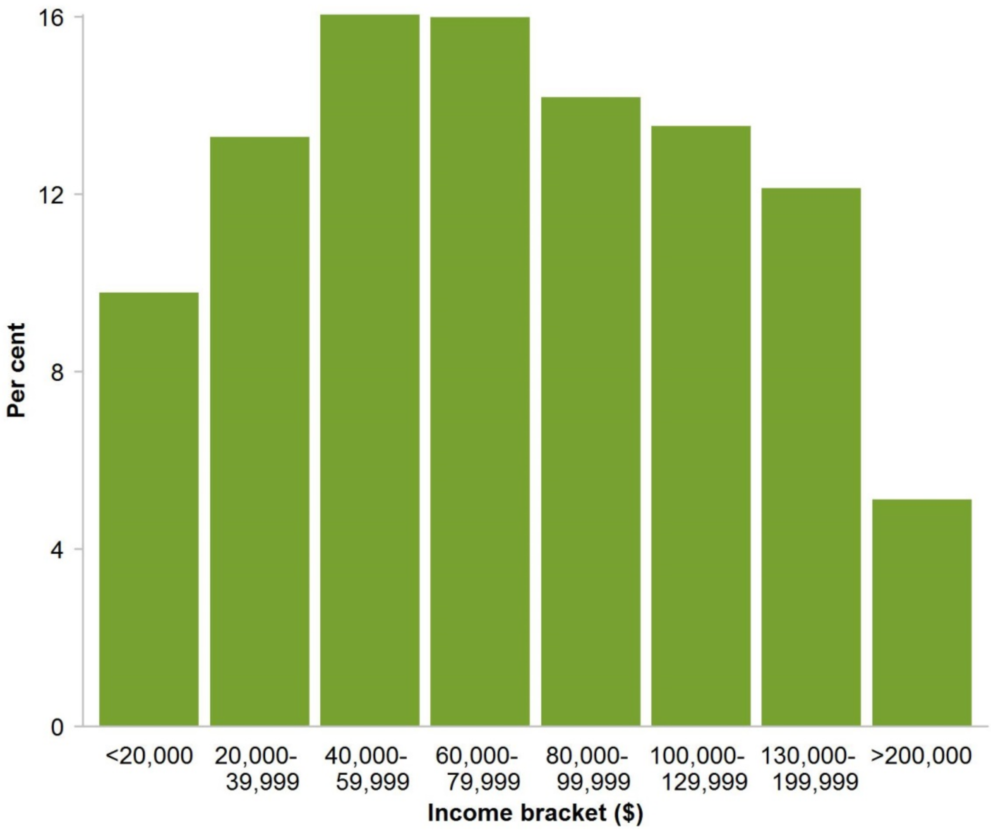
### Household income bracket

Respondents were asked about which household income bracket they were in (figure 7). The sample is broadly reflective of the income distribution of Australian households.

For some analyses, respondents were assigned to household income values based on the midway point of their income bracket, and households in the highest income bracket were assigned to an income of $250 000. This measure is used to assess features of the data and not for estimation of quantities of interest.

Figure 7 **Sample profile by household income bracket[a,b]**

N = 1996




**a** There were also 352 non-responses as this question was not mandatory.

## Financial and superannuation literacy

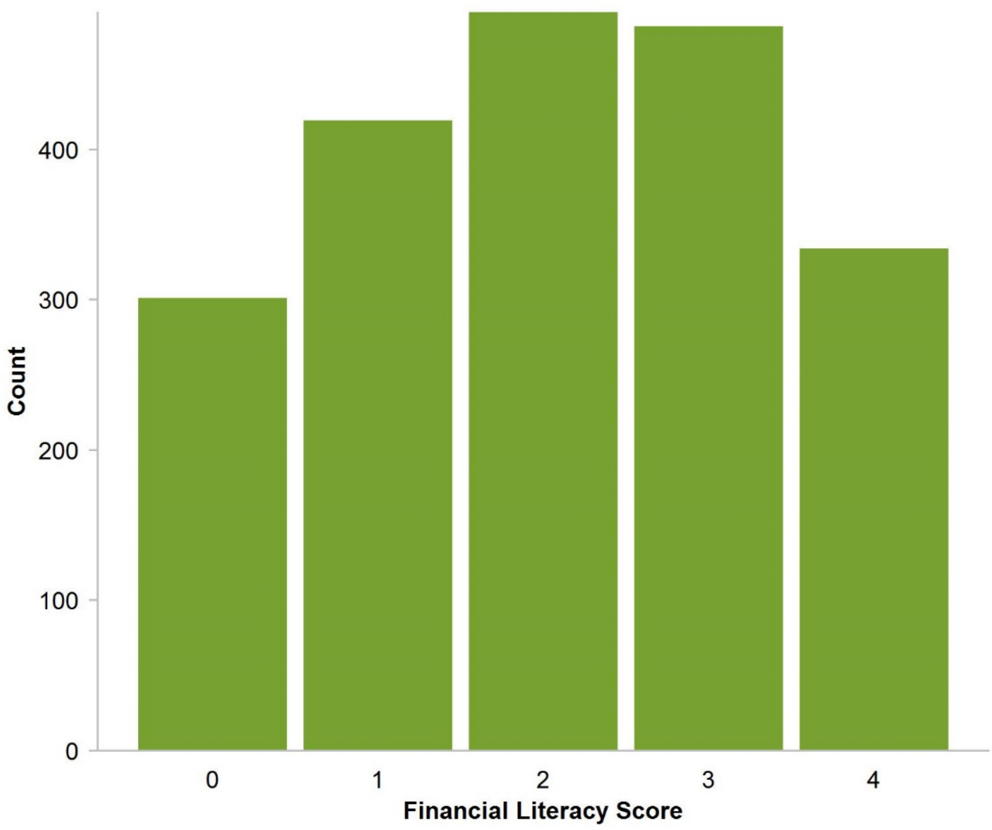
An essential observable in the context of superannuation decision making is the financial literacy and superannuation awareness of survey respondents. A measure of financial and superannuation literacy was constructed (hereafter called a financial literacy score) using the four financial and superannuation multiple choice questions in the survey. The measure is the number of correct answers by the respondent (figure 8). Scores were not calculated for respondents who had dropped out of the survey before completing all four questions.

The measure was intended to provide clear tests of selected financial literacy concepts. Collectively the four questions address basic financial concepts of compounding and diversification, and awareness of superannuation.

Despite the limited scope, the measure is still important and useful. Regression analysis (annex A.3) showed that the financial literacy score was strongly correlated with the household income of respondents. In particular, each additional point on the financial literacy score was associated with an increase of $19 000 in household income on average. This correlation was also statistically significant.

Figure 8 **Sample profile by financial literacy[a]**

N = 2027




[a] There were also 321 respondents who did not answer all four financial literacy questions.

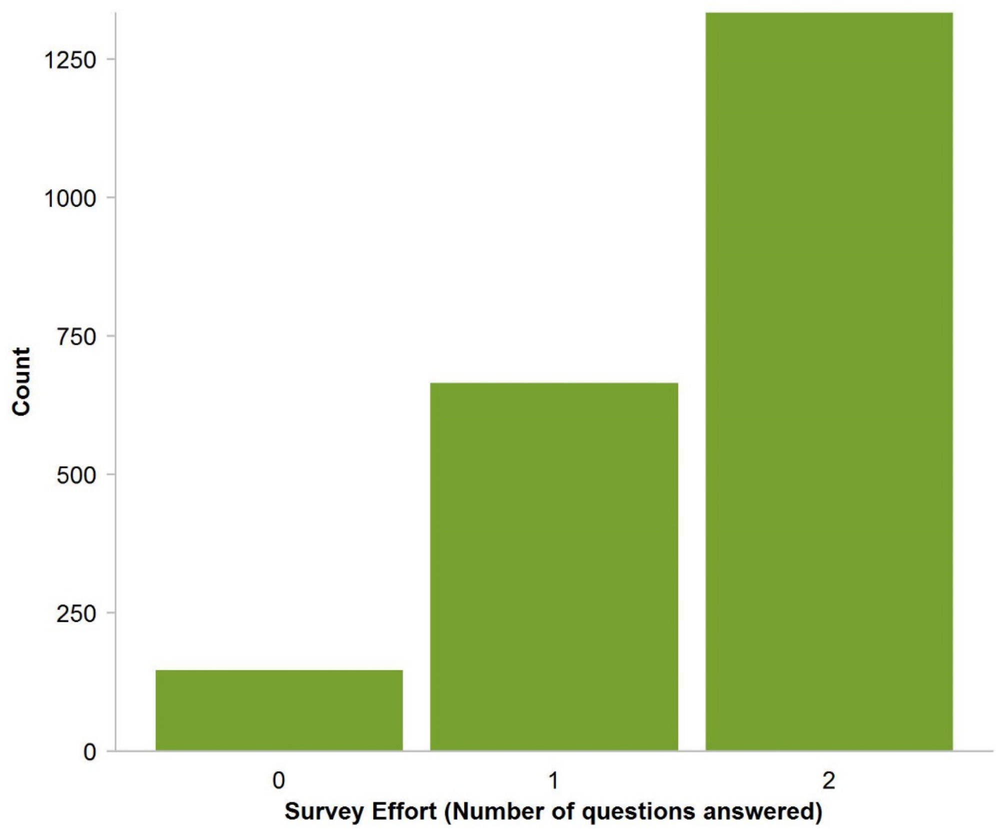
## Survey effort

The effort that respondents put into the survey is essential for the success of the analysis and understanding the results. Measuring survey effort is a difficult task. Using time taken during the choice experiment might be an intuitive choice. However, respondents may be preoccupied and complete the survey over an extended period of time. Thus it would be unclear what such a measure would represent.

The Commission constructed a measure of survey effort based on the number of open-ended response questions respondents answered *in the choice experiment section of the survey* (figure 9). This measure included two questions and a question was considered answered if a discernible response was given. The first asked why respondents made the choices they did in the choice experiment, and the second asked about the difficulty in making the choice. For most analyses it will make more sense to treat survey effort as an indicator, distinguishing between junk (those who have not answered any of the questions, or provided only undiscernible responses) and usable responses (those who have provided at least one discernible response) or respondents exhibiting 'sufficient' effort.

Figure 9 **Sample profile by survey effort[a]**

N = 2143




[a] 205 respondents dropped out before they participated in the choice experiment.

This is a reasonable measure of survey effort. As long as the respondent thought at all about the decisions they made, they would attempt to explain their decision by answering these two open-ended response questions.

Similar to the financial literacy score, the measure for survey effort is rough and provides relatively little scope for delineation between respondents. Despite this, it still appears to be a useful measure. Annex A.4 considers if the measure is adequate with some additional analysis. For brevity, the Commission will often refer to those who failed to answer any questions as exhibiting 'insufficient' survey effort and those who have answered at least one question a 'sufficient' survey effort.

## Open-ended responses

The responses to open-ended questions were almost perfectly distinct, and varied in level of detail and relevance.

To process the responses the Commission undertook an extensive manual inspection and categorisation process. The process and the final classifications of responses were internally reviewed to minimise the subjectivity in interpretation and to strike a good balance between how many categories were used (granularity) and how the categorisations lend to potential insights (tractability). Detail on this process can be found in the documentation accompanying the dataset.

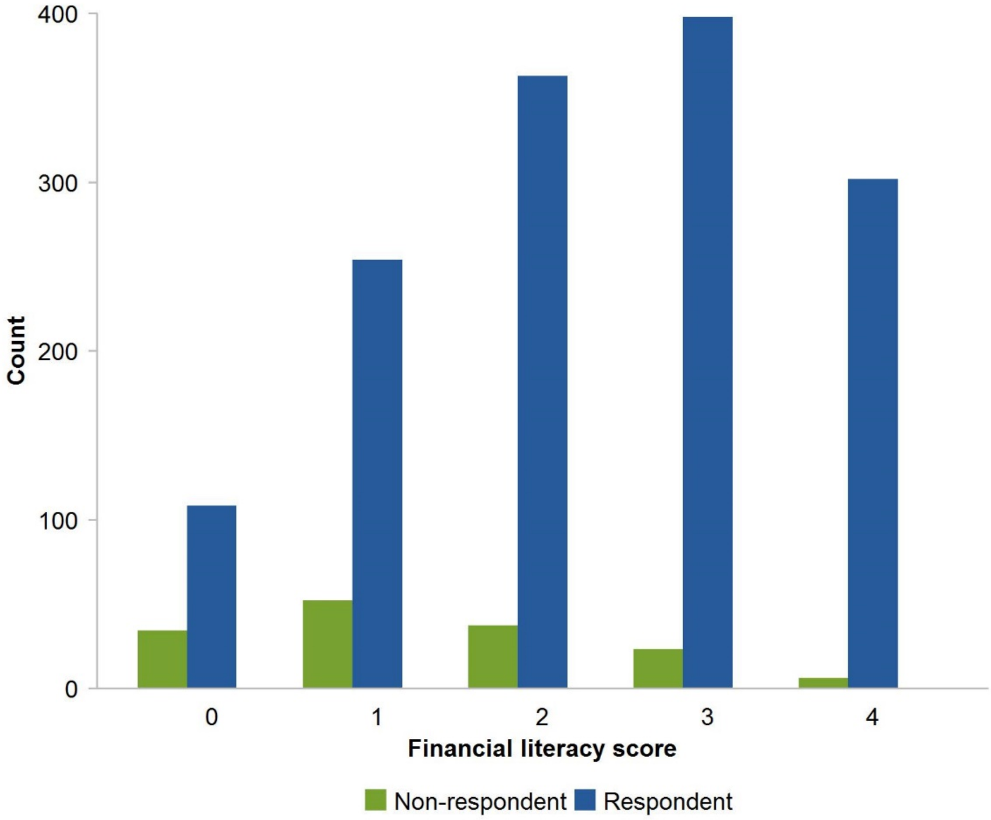
### Trade-offs in using open versus closed response data for analysis

In some cases, the survey asks questions in both a closed- and open-ended form. By comparing the two, the trade-offs and characteristics of using closed- and open-ended response data can be explored. Open-ended response data are useful because it allows respondents to submit responses free from any biases that might arise due to the framing of the options provided. It also allows for possibilities that might not have been anticipated.

On the other hand, open-ended responses also have a number of disadvantages. Open-ended responses usually require more effort and often more knowledge from the respondent. Plausibly, for this reason open-ended responses see a systematic lack of meaningful responses from younger and less financially literate respondents (the latter is shown in figure 10). This means that analysis based on open-ended responses are likely to reflect those of an older and more financially literate population, rather than the whole population.

Figure 10 **Financial literacy by respondent/non-respondent[a,b]**

N = 2027




[a] Respondents and non-respondents are those who provided a discernible response (did not provide a response) when asked about how they chose their current fund, and who answered the close-ended version of the question. [b] There were 321 respondents who did not answer all four financial literacy questions.

The other issue from open-ended responses is that there is more scope for respondents to interpret the question differently from survey intentions or fail to clearly communicate their responses. This has resulted in a lack of precision in some responses and potential misclassifications of the intentions of the respondents. This can be seen from the internal inconsistencies between a respondent's close-ended and open-ended responses. For example, 19 per cent of respondents (39 respondents) gave open-ended responses which indicated that they selected their current fund based on recommendations from others, but then also responded that they made their own independent selection in the close-ended responses.

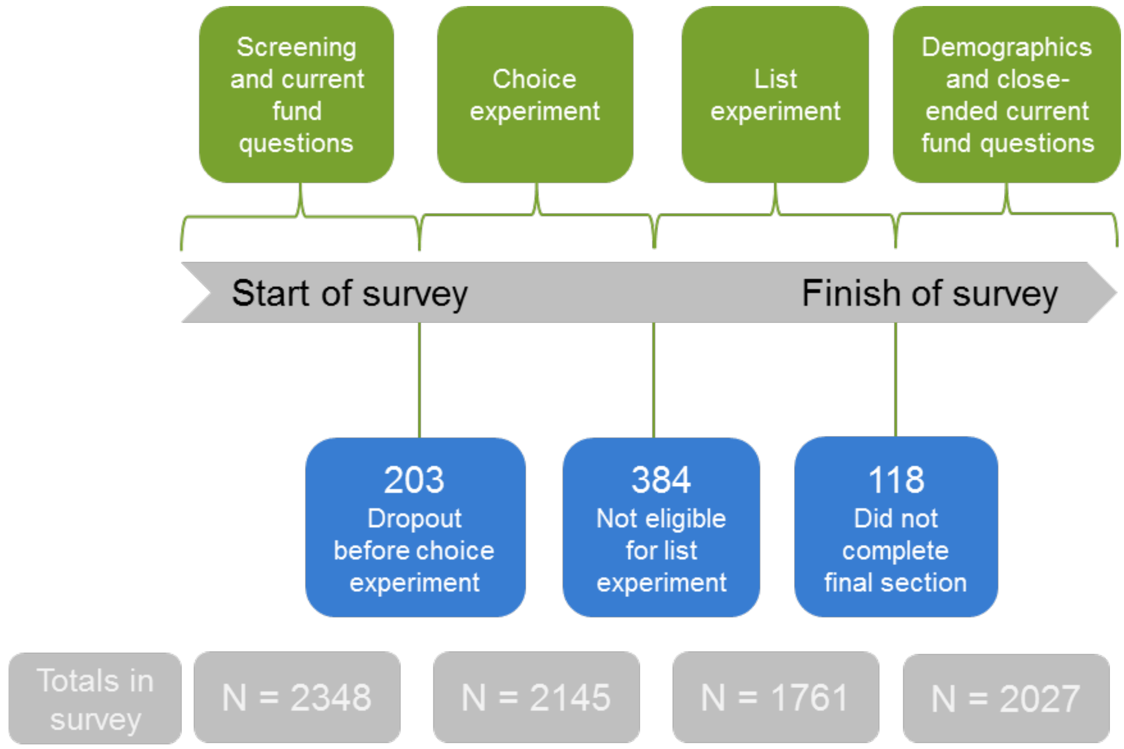
The presence of this trade-off means that when conducting analysis and when confronted with the choice, the Commission has chosen the type of response to suit the analysis at hand.

## Attrition

The majority of survey participants fully completed the survey. Different sections of the survey had different completion rates (figure 11). In the case of the list experiment, 384 participants were ineligible to participate because the questions asked participants to consider the last time they chose a super fund and these participants had never had a super fund. Such participants were skipped onto the final section of the survey.

Respondents that dropped out before the choice experiment and respondents that did not complete the final section do not appear to be from a particular demographic. Respondents that did not participate in the list experiment are systematically younger and less financially literate, likely reflecting their never having had a super fund.

Figure 11 **Participation in the survey**




## Suitability of data: balance, power and representativeness

This section provides an overview of the Commission's assessment on the suitability of data for analysis. Annex A.5 provides a full discussion of the issues, along with a brief review of causal inference concepts. In the context of this work, there are three concepts (balance, power and representativeness) relevant for a more credible analysis:

- did the randomisation result in comparable groups? Are the control and treatment groups as **balanced** as they can be? Are there steps that can be taken to improve the balance and minimise selection bias?
- considering the analyses methods, are the sample sizes large enough?
- are the experiments and analyses sufficiently **powered**?
- are the samples **representative** of members?

Regarding balance, the Commission applied a matching technique to improve the balance in the datasets. The details of the technique and the effect on balance are detailed in annex A.5.2.

To address power and sample sizes, power simulations were conducted for the experimental groups (annex A.5.3). The Commission concluded that the choice experiment was adequately powered, but that many groups in the list experiment were relatively low powered. With regards to the list experiment this requires additional caution on drawing insights about the wider Australian population, but does not preclude the Commission from drawing conclusions on the respondents *participating in this survey.*

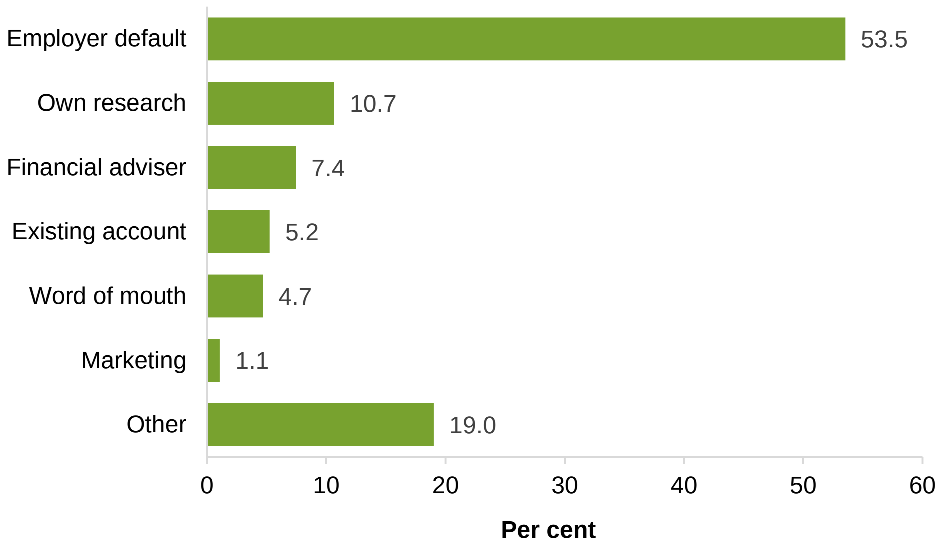
Finally, representativeness has already been discussed in earlier sections. The Commission further investigated the representativeness of the data for young Australians in annex A.5.4 and is satisfied with the representativeness for younger Australians.

# 3 How do people choose their superannuation product?

Respondents were asked what they did last time they chose a superannuation fund in two separate ways. The first provided an open-response text box option (figure 12), and the second provided a series of prompted-response options along with an ‘other’ text box option (figure 13).

Figure 12 **Fund selection method**[a,b,c,d]

Open response, n = 1773



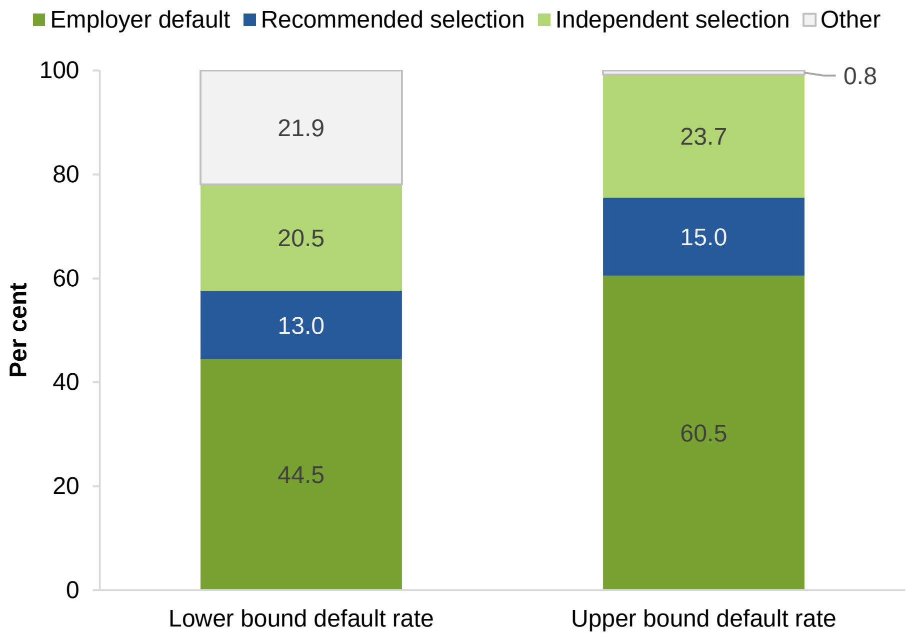

**a** Respondents were asked: 'Can you please tell us, the last time you had to select a super fund, what did you do? How did you go about it'? **b** Excludes respondents who entered a blank or invalid response to this question. Note that these respondents were younger and less financially literate on average. **c** Percentages do not sum to 100 per cent as 3.4 per cent of respondents nominated more than one reason. **d** 'Financial adviser' includes financial planners, accountants and brokers. 'Word of mouth' includes recommendations from friends, family or colleagues. 'Marketing' includes workplace seminars. 'Other' indicates any other reason, including unclear or incomplete responses, and individuals who said they have never had to make a decision or could not recall.

The estimates produced by both the open- and closed-responses are below most existing estimates of default rates (for example, 69 per cent by Colmar Brunton (2010), and 70 per cent by the Grattan Institute (2014) imputed from ABS (2009)). This is may be due to the larger set of options respondents had to choose from in the Commission's survey. In the closed environment, there was a possibility of some individuals perceiving their use of their employer's default fund as a recommended or independent selection.[7] This kind of re-casting may be even more likely for open responses. Having such a set of options differs from the research cited above, which used binary questions about whether a respondent used their employer's default fund or not.

7 As the closed-ended questions came after the choice experiment, the Commission has tested if being in the assisted choice group had any influence on respondents in their close-ended nomination questions. The Commission found no evidence to support such a concern (annex A.7.1).

Figure 13 **Fund selection method**[a,b,c,d,e]

Closed response, n=2008




[a] Respondents were asked 'Can you please tell us, the last time you had to select a super fund, what did you do? How did you go about it'? A menu of options was presented, and an 'other' open-response box was also included. Four categories produced by this 'other' option received less than one per cent each and are therefore not included individually. [b] 'Employer default' pertains to the response option 'I used my employer's default fund'; 'independent selection' to 'I made my own selection independent of anyone else'; and 'recommended selection' to 'I selected a fund recommended by someone else'. [c] The lower bound involved putting those who have never had to make a decision and those who could not recall in the 'other' category. The sample size was 2008. [d] The upper bound estimate involved removing those who have never had to make a decision from the sample, and assuming those who could not recall went with their employer's default fund. [e] In both sets of estimates 'other' included the 14 responses that came through the open-response 'other' box.

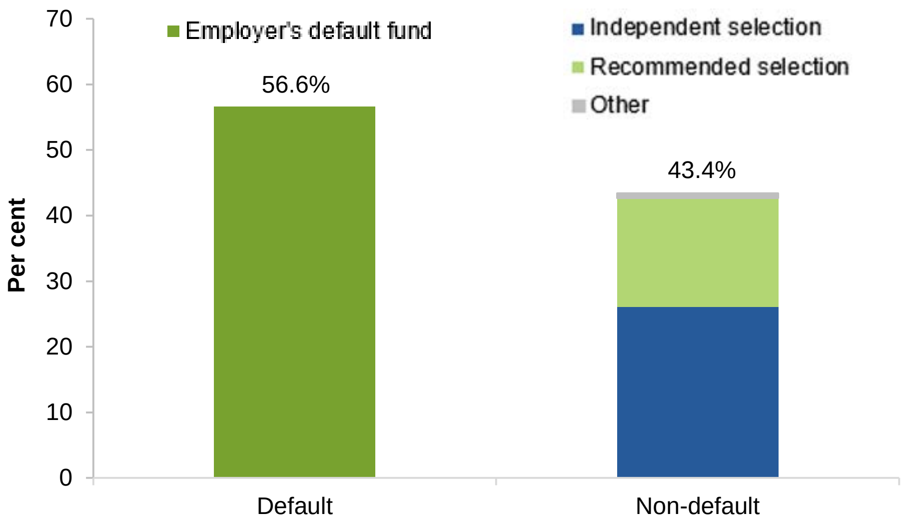
## Who defaults?

A probit regression was estimated to test the relationship between demographics and the probability of defaulting (complete results can be found in annex A.6).

The dependent variable is built from the closed-response question: 'Can you please tell us, the last time you had to select a super fund, what did you do? How did you go about it?'. The closed-response data was used as the open-response data had a large portion of invalid answers, which were systematically related to various demographic variables. Individuals who defaulted were defined as those who selected 'I used my employer's default fund', and non-defaulting respondents were defined as those who selected 'I made my own decision independent of anyone else', and 'I selected a fund recommended by someone else'. It also included two low-frequency responses that came through the open-ended ‘other’ option that included ‘Used an SMSF’ and ‘Used my existing fund’. Respondents who stated that they had never been in a decision-making situation or could not recall what they did were excluded from the sample. The final sample size was 1570 (figure 14).[8]

Figure 14 **Default rates[a]**

N = 1570




[a] Other includes ‘Used an SMSF’ and ‘I used my existing fund’.

The explanatory variables included gender, age bracket, financial literacy score, the highest level of education attained, and household income bracket. All of these variables were included in indicator form.

It is important to note that there may be a timing inconsistency between the dependent variable and some independent variables. For example, the age of the respondent at the time of the survey is likely to be different than their age when they last made a decision on a superannuation fund. The same logic could apply to education and income. This means the relationship between the dependent and these explanatory variables could be understated. For example, it is reasonable to posit that there might be an underlying positive relationship between how old an individual is when they have to make a decision, and that decision being a default selection. However, with a respondent’s current age likely to be higher than it was at this decision point, the data cannot accurately capture that underlying relationship.

Upon estimation, the probit regression model showed that current demographics explain very little of the variation in default rates in the data, with a pseudo R-squared of 0.0284.[9]

8 This sample is smaller, and therefore the default rate different, than figures 12 and 13 as only observations that had completed all demographic questions were included.

The few coefficients that have both economic and statistical significance relate to gender, and to higher levels of education and household income. With the other variables held constant, the model predicts that: [10]

- a female is 7.9 per cent more likely to default than a male
- someone with a postgraduate education is 12.2 per cent less likely to default relative to someone with a less than year 12 education
- individuals from households with income in the $60 000 to $80 000, $80 000 to $100 000, and $130 000 to $200 000 brackets are all about 16.5 per cent less likely to default, compared with someone with a household income less than $20 000. Further, those with a household income above $200 000 are 25.5 per cent less likely to default.

The estimated coefficients can also be used to estimate the probability of defaulting for a set of hypothetical individuals. These individuals include two potential new workforce entrants (given the proposed switch to the 'first-timer' pool), and two older individuals for comparison purposes (table 1). As can be observed, the small amount of variation in default rates explained by demographic data mostly stems from variation in education and household income brackets (figure 15).

Table 1 **Hypothetical individuals**

| | *New workforce entrant 1* | *New workforce entrant 2* | *Older individual 1* | *Older individual 2* |
|---|---|---|---|---|
| Gender | Male | Female | Male | Female |
| Age bracket | 15 – 24[a] | 15 – 24[a] | 35 – 44 | 45 – 55 |
| Fin. literacy score | 1 | 3 | 2 | 4 |
| Education level | Year 12 | Bachelor | Diploma | Postgraduate |
| Household income bracket | $20 000 – $40 000[b] | $40 000 – $60 000[c] | $60 000 – $80 000 | $130 000 – $200 000 |

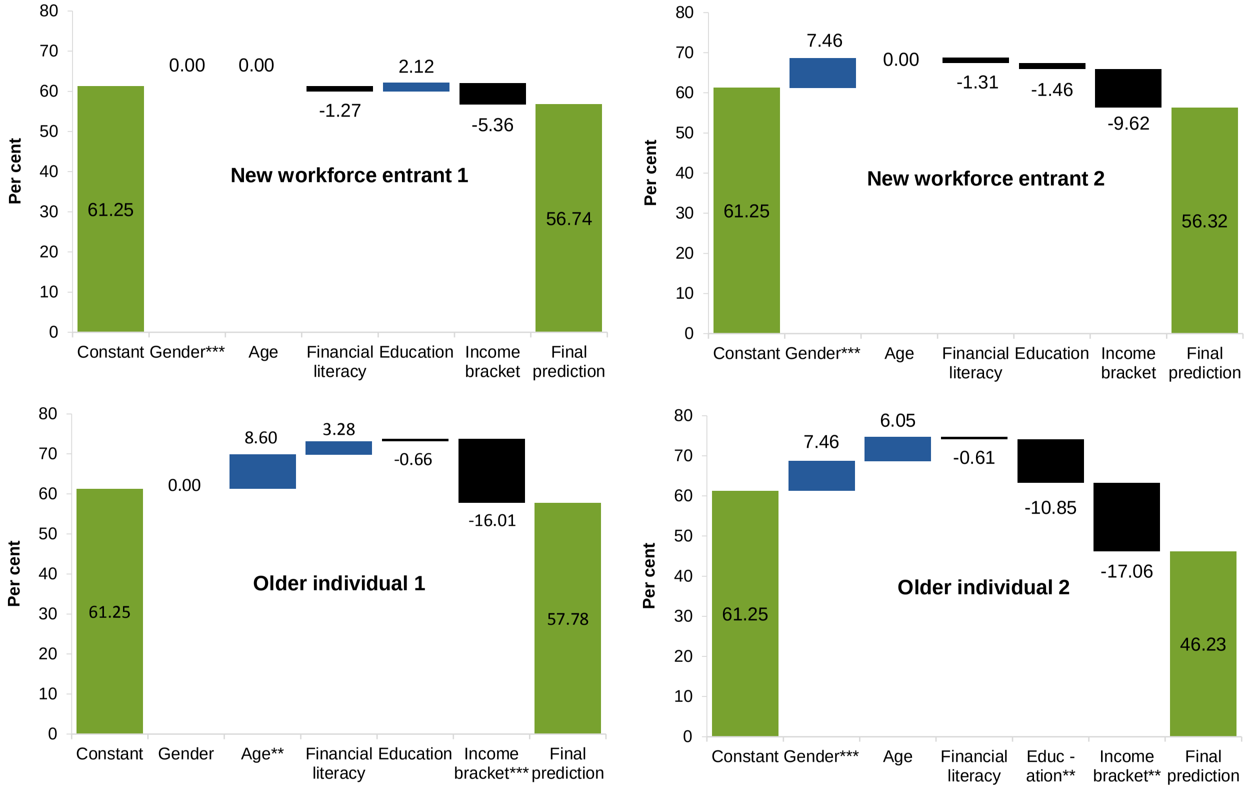
**a** 60 per cent of new entrants to the superannuation system are under 25 (PC 2017). **b** The average income for a new workforce entrant is $21 000 (PC 2017). Using this figure for household income essentially assumes the individual has moved out of the family home. **c** In 2015, the median salary for bachelor degree graduates aged less than 25 in their first full-time employment was $54 000 (GCA 2015). As before, using this for household income bracket essentially assumes the individual has moved out of the family home.

9 This means that variation in the explanatory variables can explain 2.84 per cent of the variation in the default rate.

10 These predicted marginal effects are computed by holding indicator variables in other characteristic categories at their sample average and projecting the entire equation onto the standard normal distribution. All of the examples used were statistically significant at the 1 per cent level, except the postgraduate effect, which was statistically significant at the 5 per cent level.

Figure 15 **Probit model results[a,b,c]**

For a set of hypothetical individuals (table 1)




**a** The 'constant' value pertains to the model's constant term. Given the use of indicator variables this represents the predicted probability of defaulting for a male, aged between 15 – 24, with a financial literacy score of zero, a below year 12 education level, and a household income less than $20 000. **b** Each incremental change represents the marginal effect of the individual's given characteristic on the probability of defaulting, and the 'end' value is the actual predicted probability of defaulting for the individual. **c** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively.

# 4 Attitudes towards super

## List experiment results

As explained earlier, a list experiment is a technique used to test sensitive statements that respondents might otherwise be hesitant answering truthfully (section 1). Regression analysis was used to estimate the proportion of respondents that agree with each of the statements. The statements are broken down into two broad categories: those that relate to fund selection engagement and understanding (table 2), and those that relate to broader system engagement and understanding (table 3). The regressions were run using a trimmed sample to address imbalance concerns in the 'raw' dataset (that is, the treatment and control groups looked different demographically). The full results of the estimation (for both the 'raw' and 'matched' sample) can be found in annex A.2.3, and details regarding the matching process that trimmed the sample can be found in annex A.5.2.

Table 2 **List experiment results[a]**
Fund selection engagement and understanding

| *Sensitive question* | *Per cent agree* | *Std. error* | *p-Value* |
|---|---|---|---|
| **Significant results[b]** | | | |
| I didn't really know what sorts of information I should consider in making my choice | 63.5 | 18.8 | 0.0011 |
| I trusted other people to make the decision for me | 52.0 | 17.5 | 0.0035 |
| I was influenced by funds' advertising | 46.7 | 16.7 | 0.0060 |
| I felt overwhelmed by the number of choices before me | 36.7 | 19.5 | 0.0626 |
| I didn't end up making a fund choice at all | 36.7 | 18.1 | 0.0454 |
| I was actually annoyed at having to choose a fund | 36.7 | 19.6 | 0.0633 |
| I felt overwhelmed by the amount of information I was supposed to consider in making my choice | 32.7 | 19.4 | 0.0954 |
| **Insignificant results[b]** | | | |
| I pretended to care about the decision, but I didn't really care at all | 29.9 | 19.5 | 0.1281 |
| I wished my employer had just recommended a fund suitable for my lifestage and circumstances | 28.3 | 18.9 | 0.1366 |
| I went with a super fund I already had, without considering any other information | 28.2 | 17.7 | 0.1142 |
| I was annoyed at how much time and energy it took to choose a fund | 25.5 | 18.4 | 0.1682 |
| I chose a fund that sounded familiar to me | 24.6 | 20.3 | 0.2273 |
| I chose the first fund I came across | 22.5 | 20.2 | 0.2676 |
| I chose a fund at random | 15.1 | 17.3 | 0.3856 |

**a** Note that recipients were asked to reflect on the last time they had to select a superannuation fund. **b** Significance was categorised as a combination of statistical and economic significance, as they tended to coincide.

Table 3 **List experiment results**[a]

System engagement and understanding

| *Sensitive question* | *Per cent agree* | *Std. error* | *p-Value* |
|---|---|---|---|
| **Significant results**[b] | | | |
| I didn't fully understand who was supposed to make super "contributions", or how | 56.0 | 20.8 | 0.0085 |
| I didn't fully understand how the super contributions I made now would affect my retirement income later | 54.5 | 21.0 | 0.0107 |
| I had a good understanding of what a fund's "asset allocation" referred to | 47.1 | 20.3 | 0.0228 |
| I didn't really know what it meant to say a fund had a certain level of "risk" | 41.7 | 17.3 | 0.0175 |
| I didn't really know what super was for | 41.4 | 21.7 | 0.0596 |
| I really gave no thought to how my choices might affect my account balance | 35.4 | 16.6 | 0.0348 |
| **Insignificant results**[b] | | | |
| I didn't really know what a "super fund" was | 24.6 | 20.2 | 0.2277 |
| I didn't understand the meaning of super "returns" and how they were calculated | 24.4 | 24.0 | 0.3111 |
| I couldn't care less about super | 21.1 | 16.4 | 0.2006 |
| I felt uneasy thinking about my retirement | 15.1 | 17.5 | 0.3901 |

**a** Note that recipients were asked to reflect on the last time they had to select a superannuation fund. **b** Significance was categorised as a combination of statistical and economic significance, as they tended to coincide.

As can be observed, the list experiment produced a mix of significant (both economic and statistical) and insignificant results. The Commission has not tried to reconcile these results with others from the list experiment or the survey more broadly.

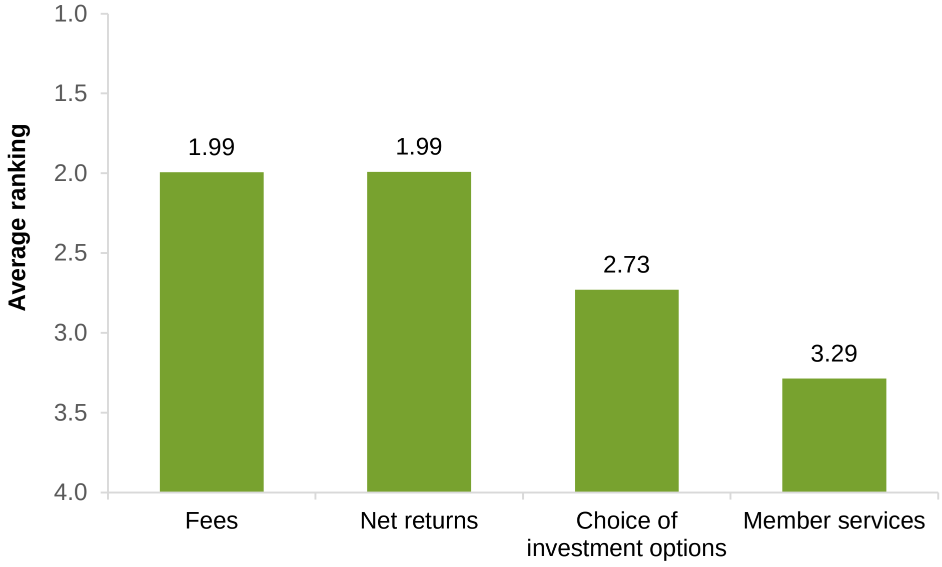
## What product features are important to people?

Respondents were asked to rank four typical superannuation product features in order of importance when selecting a superannuation fund. These were: fees, net returns, the choice of investment options, and member services. It is important to note that these four product features are not exhaustive in representing a superannuation product. Furthermore, the member services category is broad and may mean different things to different people.

On average, fees and net returns were deemed the most important factors (figure 16). About 75 per cent of respondents put either fees or net returns as the most important product feature in superannuation, and 56 per cent of respondents ranked member services as the least important product feature.[11]

Figure 16 **Product preferences[a]**

Average ranking of importance of product feature in set




[a] A response of 1 indicates most important, and 4 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

The above rankings were consistent across all the income and education levels of respondents with all categories of respondents ranking fees and net returns as more important than the choice of investment options and member services. Nonetheless, the relative importance of member services declined for older members (figure 17).

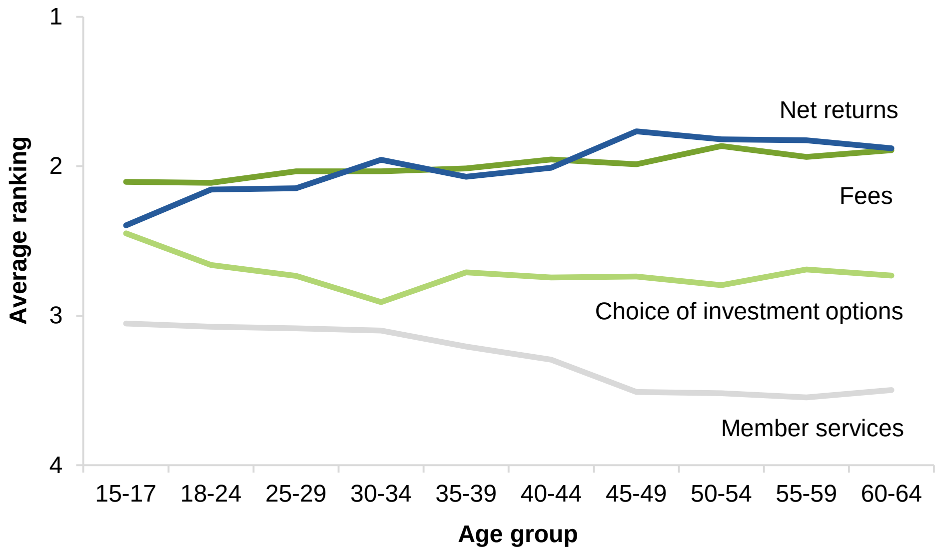
There was also a noticeable relationship between how respondents scored on the financial literacy questions, and how they tended to rank the four product features presented (figure 18). Respondents who scored higher for financial literacy tended to value net returns and fees substantially higher than member services.[12]

11 As the product preference questions came after the choice experiment, the Commission has tested if being in the assisted choice group had any influence on respondents in their product preference questions. The Commission found no evidence to support such a concern (annex A.7.2).

12 It is worth cautioning that the clustering of these rankings for low levels of financial literacy may represent randomised responses from confused or disinterested respondents rather than a genuine ordering of preferences.

Figure 17 **Product preferences by age group[a]**

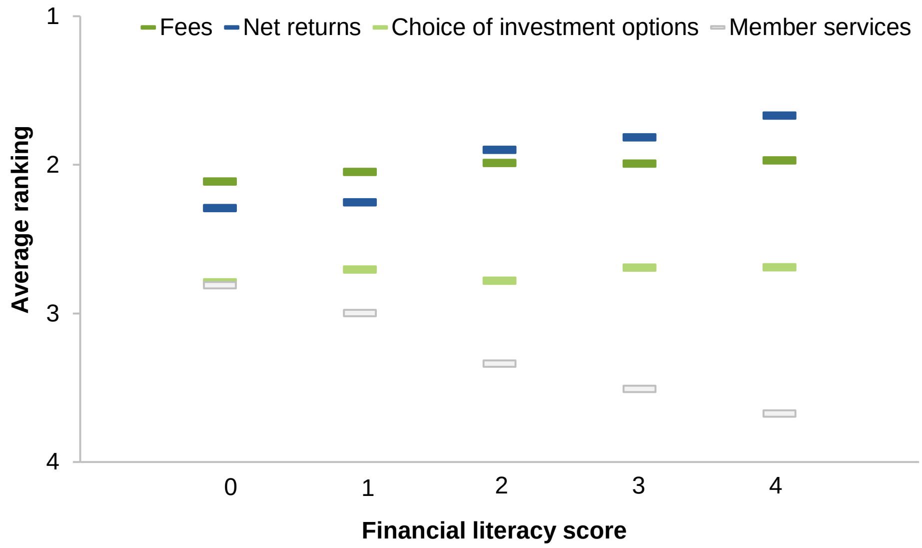
Average ranking of importance of product feature




[a] A response of 1 indicates most important, and 4 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

Figure 18 **Product preferences by financial literacy[a,b]**

Average ranking of importance of product feature




[a] Financial literacy was measured via four simple questions relating to superannuation and finance more generally. A respondent's score is the number they got correct. [b] A response of 1 indicates most important, and 4 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

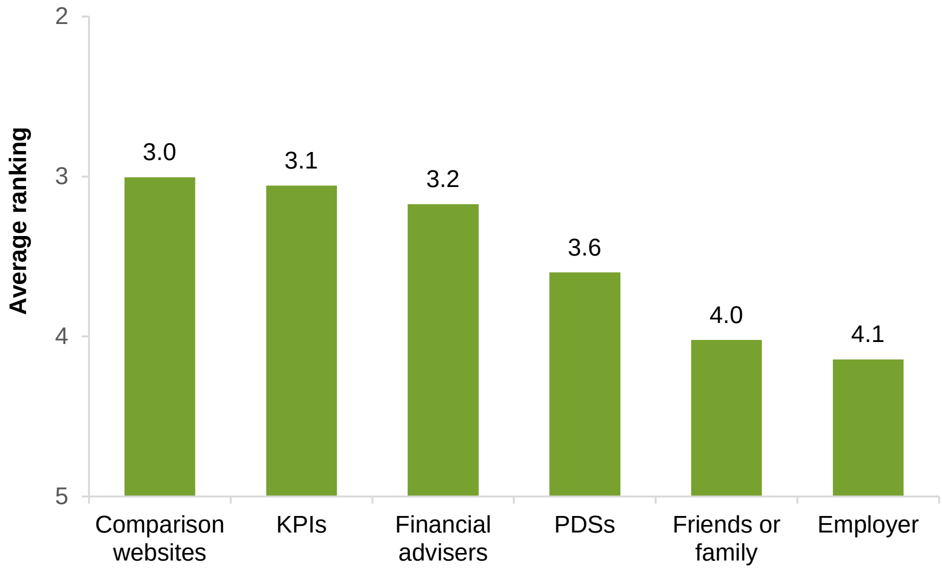
## What information sources are important to people?

Respondents were also asked to rank six potential sources of information in order of importance if they had to select a superannuation fund. Respondents tended to rank comparison websites and key performance indicators as the most important sources of information, with employers being the least important (figure 19).[13]

It is important to note that the respondents were asked this question in a prospective manner. That is, they were asked about picking a superannuation fund in the future. The Commission has not tried to reconcile any of these rankings by respondents with responses to questions about their past behaviour, as some information source categories may mean different things to different people. Further, the Commission has not tested for the effect of surrounding policies, such as the prompting to get independent financial advice on various matters.

Figure 19 **Information sources[a,b,c]**

Average ranking of important of information sources



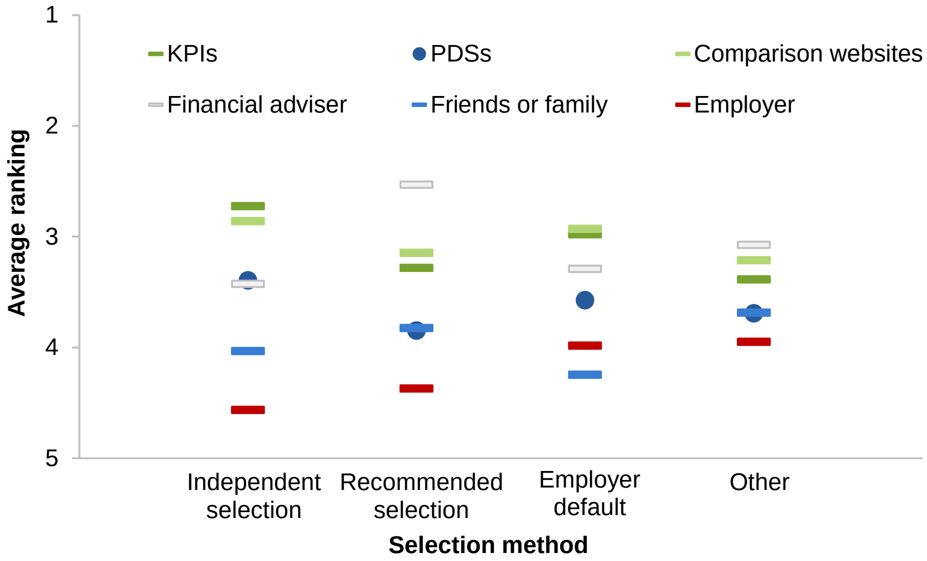

**a** A response of 1 indicates most important, and 6 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance. **b** ‘KPIs’ refers to ‘key performance indicators for a set of funds’, ‘PDSs’ refers to ‘product disclosure statements from individual funds’. **c** A response of 1 indicates most important, and 6 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

13 As the information source questions came after the choice experiment, the Commission has tested if being in the assisted choice group had any influence on respondents in their information source questions. The Commission found no evidence to support such a concern (annex A.7.3).

These average rankings were consistent across all household income and education levels. However, there were noticeable differences in average rankings by fund selection style (figure 20). Independent decision makers held key performance indicators and comparison websites as particularly important, and recommended decision makers held financial advisers as particularly important. Those who defaulted ranked employers higher than other selection style types, but still lower than all the other options except friends and family. The clustering for the group 'other' is potentially due to respondents selecting randomly (keeping in mind the other group is predominantly people who have never had a fund or could not recall how they selected their fund).

Figure 20 **Information sources by selection style[a,b]**

Average ranking of importance of information source




[a] The specifics of the selection style data is the same as in figure 13. [b] A response of 1 indicates most important, and 6 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

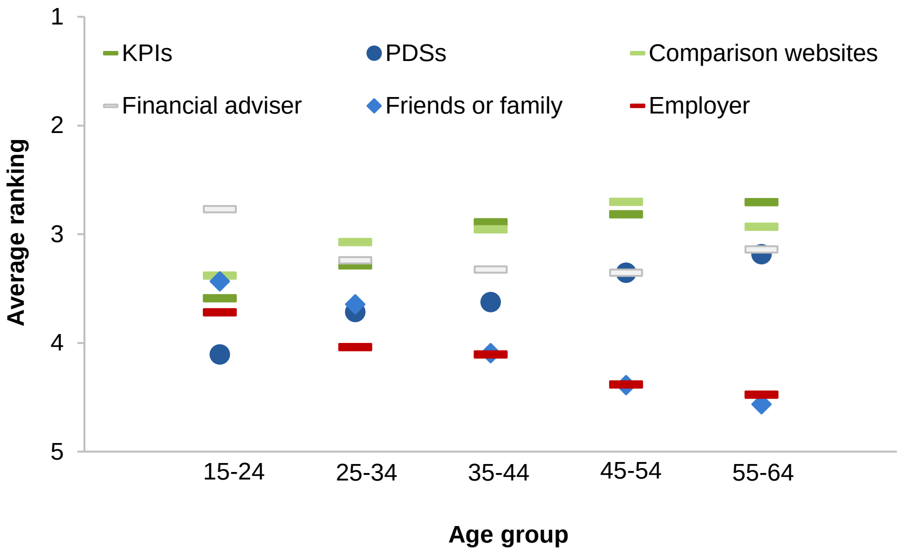
There is also a noticeable relationship between age and the average ranking of information sources (figure 21). Older individuals value KPIs, comparison websites and PDSs more so than their younger counterparts. And younger individuals seem to value input from financial advisers, employers, and friends and family more so than older age groups.

And last, the average ranking of information sources varied by financial literacy score (figure 22). As financial literacy improves, individuals place less value on their employer, and friends or family, and more value on KPIs and comparison websites.[14]

[14] As with the product feature rankings, it is important to note that the clustering at low financial literacy scores may reflect randomisation by confused or disinterested respondents rather than a genuine reflection of their preferences.

Figure 21 **Information sources by age group**[a,b]

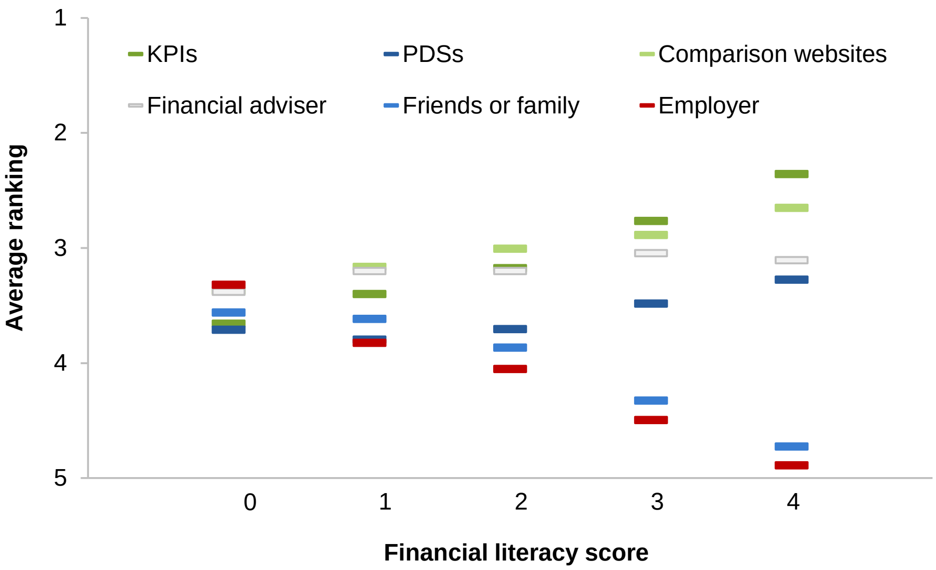
Average ranking of importance of information source




**a** A response of 1 indicates most important, and 6 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

Figure 22 **Information sources by financial literacy**[a,b]

Average ranking of importance of information source




**a** A response of 1 indicates most important, and 6 indicates least important. The vertical axis is in reverse order to illustrate that a lower average ranking indicates a higher level of importance.

# 5 Nomination rates in the choice experiment

As explained in section 1, most respondents participated in a choice experiment where they were randomly allocated into an unassisted choice group (control) and an assisted choice group (treatment) and asked to make a hypothetical nomination decision in these settings. The remainder of the supplement considers analysis about the choice experiment.

## How many nominate?

This section examines the nomination rates of respondents choosing a valid fund or product in the choice experiment. As the Commission encountered some data issues with respect to nominations by respondents (box 2), this section will start by considering regression analysis to examine the treatment effects.

### Box 2 Respondents skipping nominations

Respondents in the assisted choice task were not required to make a choice and had the option of moving on without providing a choice. This design feature corresponds to those who would not nominate and be put in the fund of last resort in the Commission's assisted employee choice model.

There were 272 respondents in the assisted choice task who were recorded as not making a choice. However, based on examining the open-ended responses provided about the choices made, a conservative estimate of 70 per cent of these represent cases where respondents actually intended on making a choice. For example, they may have simply forgotten to click on their choice and clicked next.

These 272 respondents have been excluded from the analysis, as including these respondents would require many judgment calls. The implications of this issue are discussed further in annex A.8.2. The annex also contains an additional regression that was conducted to check for robustness of the qualitative results and to get a better sense of the quantitative results.

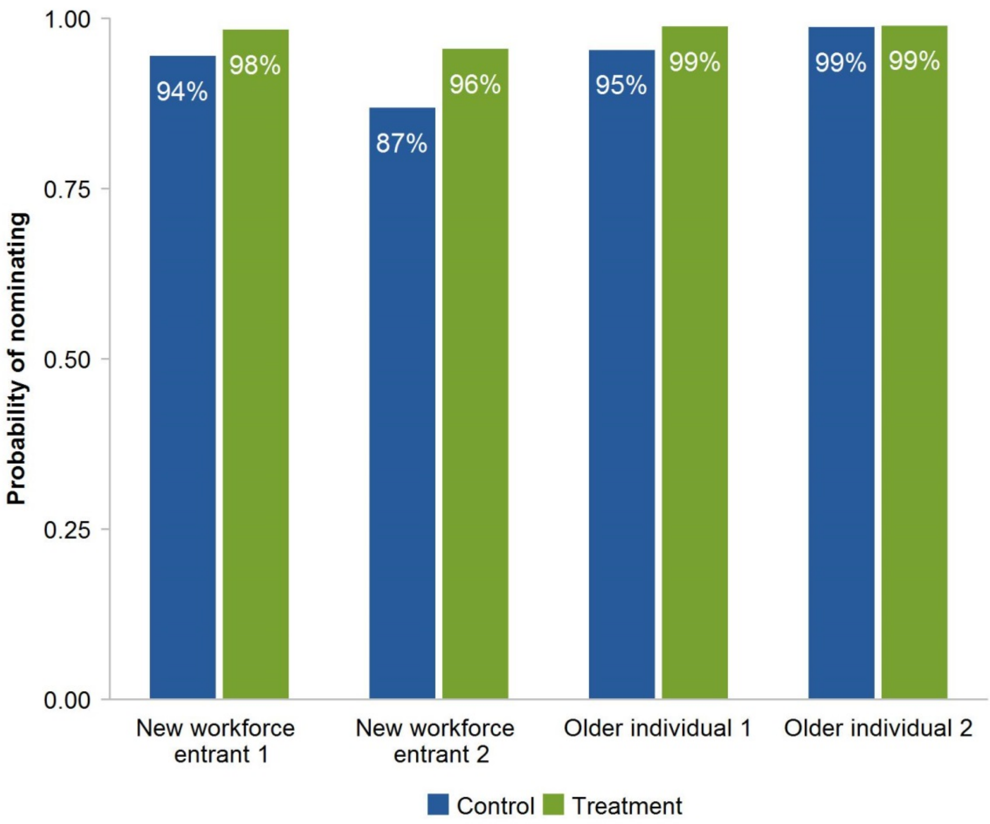
All respondents, irrespective of whether they nominated a fund or not, were presented with questions about why they made their choice and the difficulty in arriving at the choice. Thus the survey effort measure is unaffected.

The Commission conducted probit regression (with matching) to explore the treatment effects and how demographics may also affect nominations. After matching, this regression used 1069 responses. Annex A.8.1 provides the full details about this regression.

Figure 23 shows the model predictions for each of the hypothetical persons outlined in table 1. The left and right bars represent the hypothetical person in the unassisted choice group and the assisted choice group, respectively. Nomination rates are above 80 per cent across the board. Treatment effects are positive or nil in each case.

Figure 23 **Nomination rate predictions[a,b]**

Hypothetical individuals




[a] New workforce entrant two in this case has been specified with a financial literacy score of two instead of three. This is due to there being coefficients associated with a financial literacy score of three having very large standard errors. [b] Unassisted choice selections were manually processed by inspection and bucketed into categories. Cases where respondents have 'attempted' to nominate have been classified as respondents nominating. Examples of these include, where the respondent has mistakenly nominated something else such as a health insurance fund, or where the respondent has submitted a type of fund they would nominate, such as an industry fund.

There is evidence that the interaction effects between treatment and financial literacy scores are significant. In particular, treatment effects are much larger for those with lower financial literacy than for those with higher financial literacy. At the same time, these coefficients are statistically insignificant at the 10 per cent level. Despite this, annex A.8.1 reports the results of the same regression without interaction coefficients and found that there is an economically and statistically significant treatment effect. Taken together, this suggests that the assisted employee choice model improves the rate of nominations on average (and in this case specifically averaged across financial literacy scores). However more data would be required to properly identify potential varying treatment effects.

Regression results showed that having an existing fund increases the likelihood of nominating a fund. This result is economically and statistically significant, but is not substantially more important than other factors. This suggests that having an existing fund is not the only factor driving high nomination rates.

Financial literacy scores of three and four appear to have a positive effect on nomination rates. In the case of scores of three the standard errors are very large, but this is due to almost everyone in the assisted choice group with a score of three nominating. Considering the regression without interaction effects (which does not have large standard errors), scores of three increase nomination rates. This is both economically and statistically significant. On the other hand, scores of four have relatively small effect sizes and are statistically insignificant.

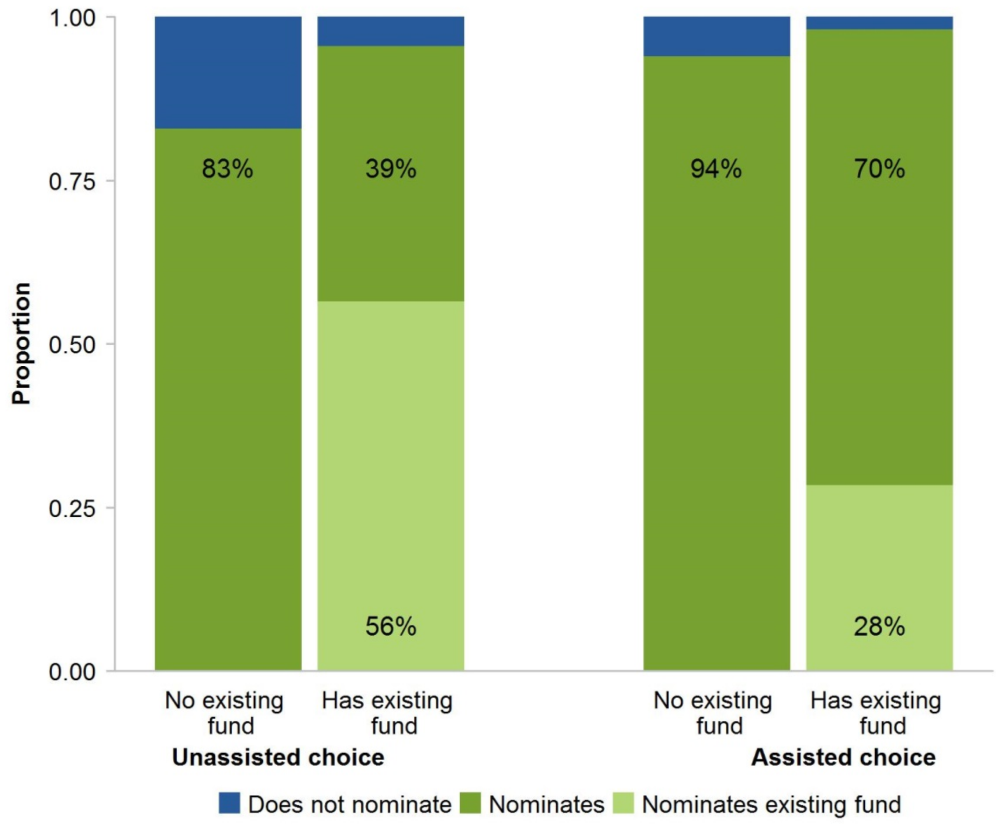
The household income bracket of respondents appeared to matter. All coefficients are large and positive with most being statistically significant, meaning that respondents with household income over $20 000 nominate more than those with household income below $20 000. The effect is strongest for the $100 000 to $130 000 bracket.

## Having an existing fund

An important factor to explore is whether having an existing fund affected decisions in assisted and unassisted choice scenarios. Figure 24 examines these issues. Consistent with the regression analysis, there are high nomination rates across the board, even when looking at those without an existing fund (and thus presumed to currently be outside of the super system). Figure 24 also demonstrates that when respondents have an existing fund, they are more likely to stay with it in the unassisted choice world compared with the assisted choice world. Overall this figure suggests that respondents will nominate regardless of whether they are assisted or not. However in an unassisted world, respondents are more likely to stick with their existing fund rather than consider other options.

Figure 24 **Nomination rates in the choice experiment[a,b]**

N = 1868




**a** In the has existing fund columns, the nominates category corresponds to those nominating a fund other than their existing fund. **b** Three respondents experienced technical problems in their surveys and their observations been removed. 205 respondents did not participate in the choice experiment. 272 observations were removed due to the skipping nomination problem (box 2).

# 6 Shortlist design

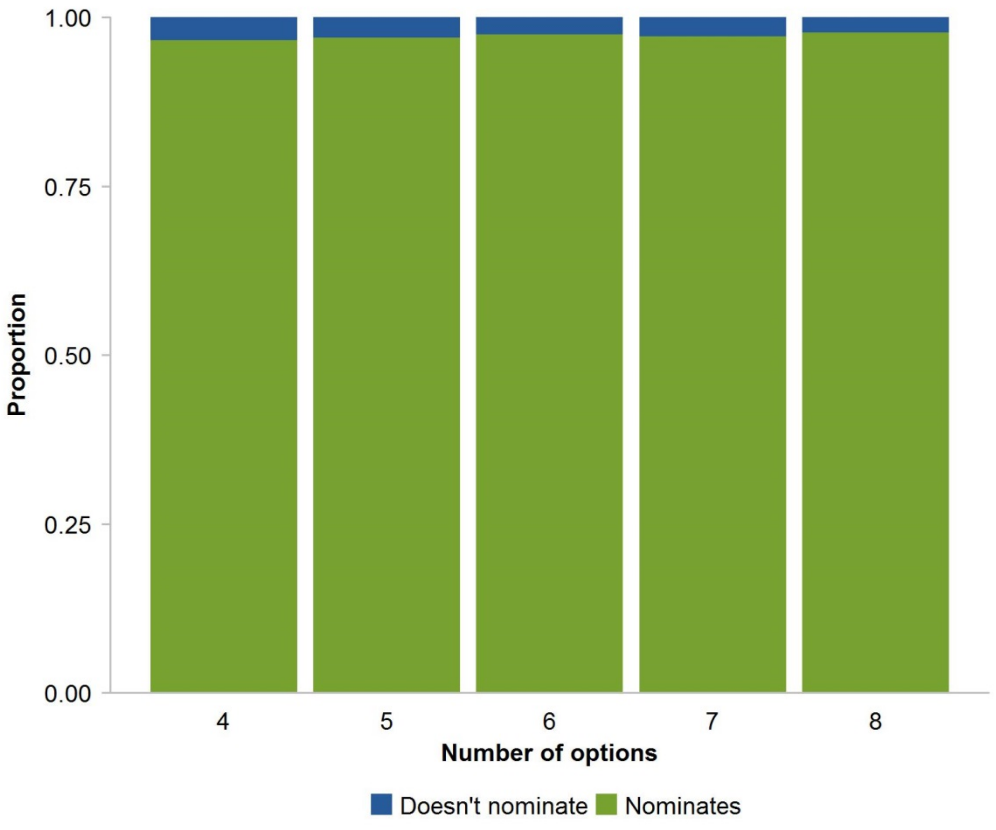
In the assisted choice group, respondents were randomly assigned to sub-groups, with each sub-group facing a different variant of the shortlist (figure 2). This section considers results from the different sub-groups to gain some insight on how best to design a shortlist. The Commission sought to test the effect of particular aspects of shortlist design on the rate of nomination of products and on the difficulty for respondents in making a choice.

## Comparing the different shortlists — nomination rates

The Commission examined whether presenting more or less options, or if presentation in terms of dollars or per cent, might improve nomination rates. Design of the shortlist based on the number of options presented (figure 25) and whether information is presented as dollars or per cent (figure 26) had almost no effect on nomination rates. Probit regression analysis was also conducted with similar results (annex A.9).

Figure 25 **Nomination rates in the assisted choice group by number of options[a,b]**

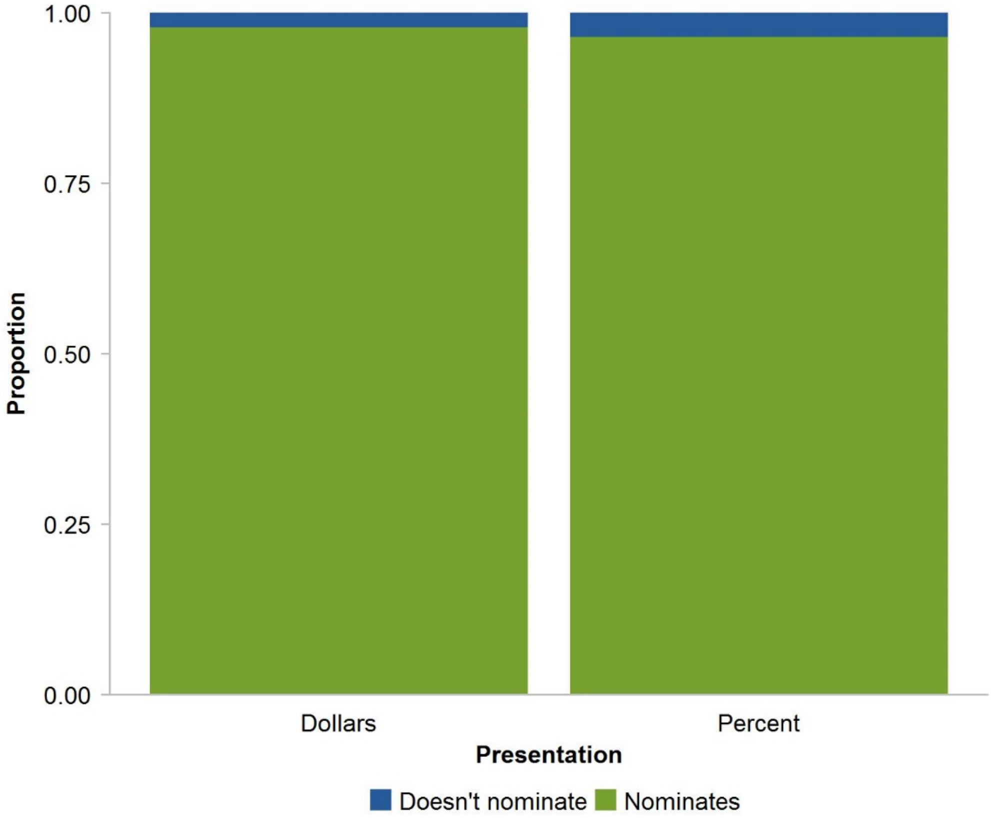
N = 1509




[a] The data used for this analysis did not include nomination skippers. Such respondents have been randomised across the sub-groups, such that the potential effects of those who genuinely intended to skip the question would be minimal. [b] Three respondents experienced technical problems in their surveys and their observations have been removed. 205 respondents did not participate in the choice experiment.

Figure 26 **Nomination rates in the assisted choice group by per cent or dollar[a,b]**

N = 1509



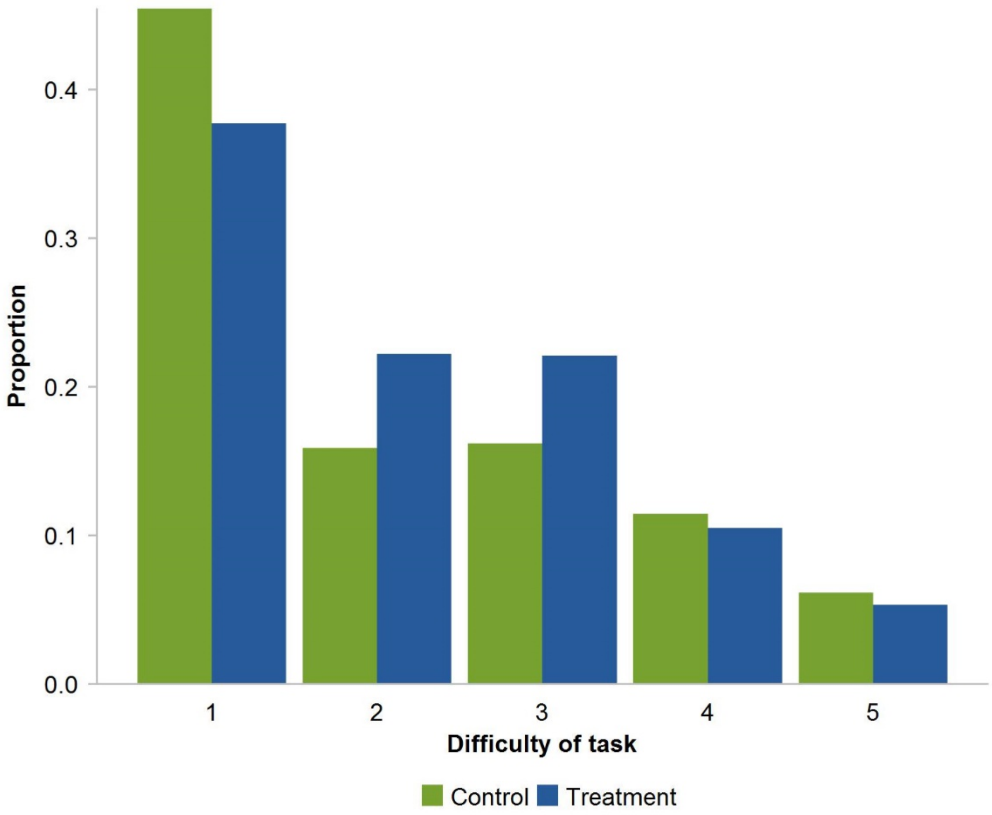

**a** The data used for this analysis did not include nomination skippers. Such respondents would have been randomised across the sub-groups, such that the potential effects of those who genuinely intended on skipping would be minimal. **b** Three respondents experienced technical problems in their surveys and their observations have been removed. 205 respondents did not participate in the choice experiment.

## Difficulty across control and treatment

The remainder of this section evaluates the difficulty experienced by respondents in making their choices. Respondents were asked to provide a difficulty score for the choice experiment, where five represents the highest difficulty.

The average difficulty score in the control group was 2.1 compared to 2.2 for the treatment group (where the treatment group consists of all assisted choice group respondents). However, these averages mask differences in the composition of the scores across the two groups (figure 27).

Figure 27 **Difficulty of choice experiment across control and treatment[a]**
N = 1866




[a] Three respondents experienced technical problems in their surveys and their observations have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem. Two respondents failed to submit a difficulty score.

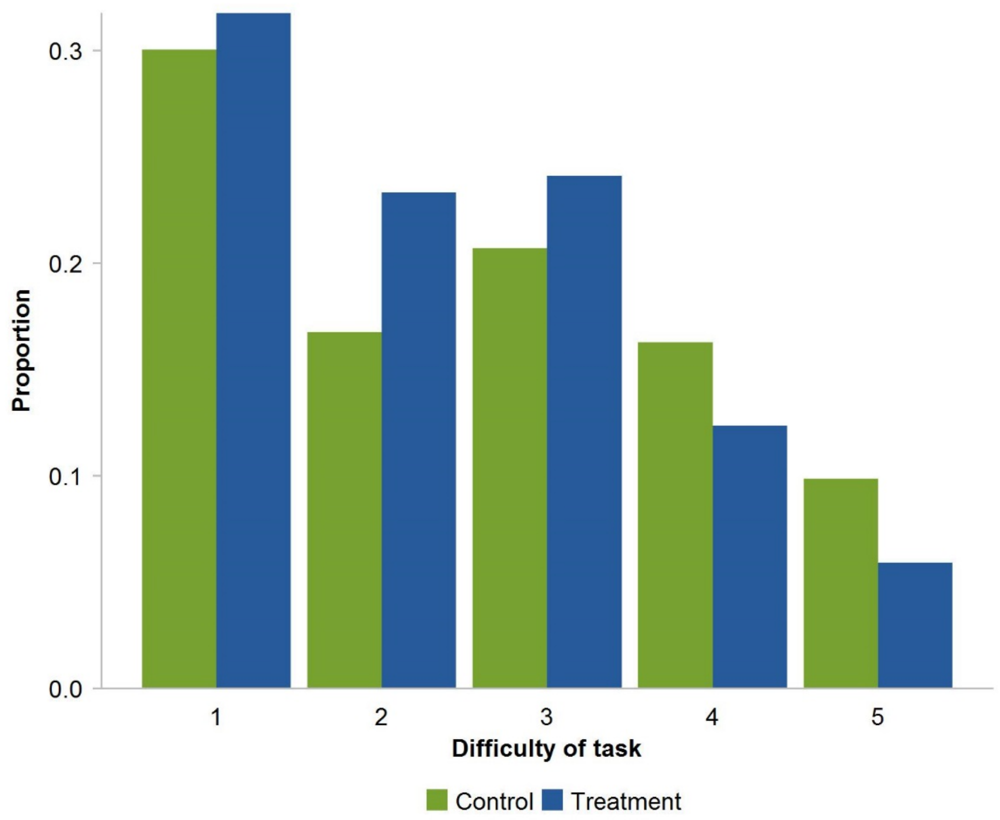
The figure shows that more unassisted choice respondents (control group) rated their task the least difficult compared with assisted choice respondents (treatment group). On the other side of the distribution, fewer assisted choice respondents rated their tasks at the higher difficulties compared with the unassisted choice respondents.

Overall this figure suggests that the unassisted choice task might be easier for respondents than the assisted choice task. As before, however it is important to consider the role of selecting an existing fund.

Figure 28 considers the role that selecting an existing fund might play in the reported difficulty of the task, by removing those who choose existing funds. The figure clearly shows in this case that the assisted choice task is easier than the unassisted choice task. There are a higher proportion of respondents rating the lower three difficulties in the assisted choice group than the unassisted choice group. On the other hand, there are substantial increases in the proportion of respondents rating difficulties as four or five in the unassisted choice groups. This suggests that many in the unassisted choice group who had existing funds may have simply nominated their existing fund, without needing to do anything more. On the other hand, in the assisted choice group, respondents may have first attempted to use the shortlist before deferring to their current fund. The figure may also reflect a 'learning effect', whereby those with existing funds have a greater knowledge of the system and are more confident in making a selection.

Figure 28 **Difficulty of choice experiments for those who did not choose existing funds**[a]

N = 1351

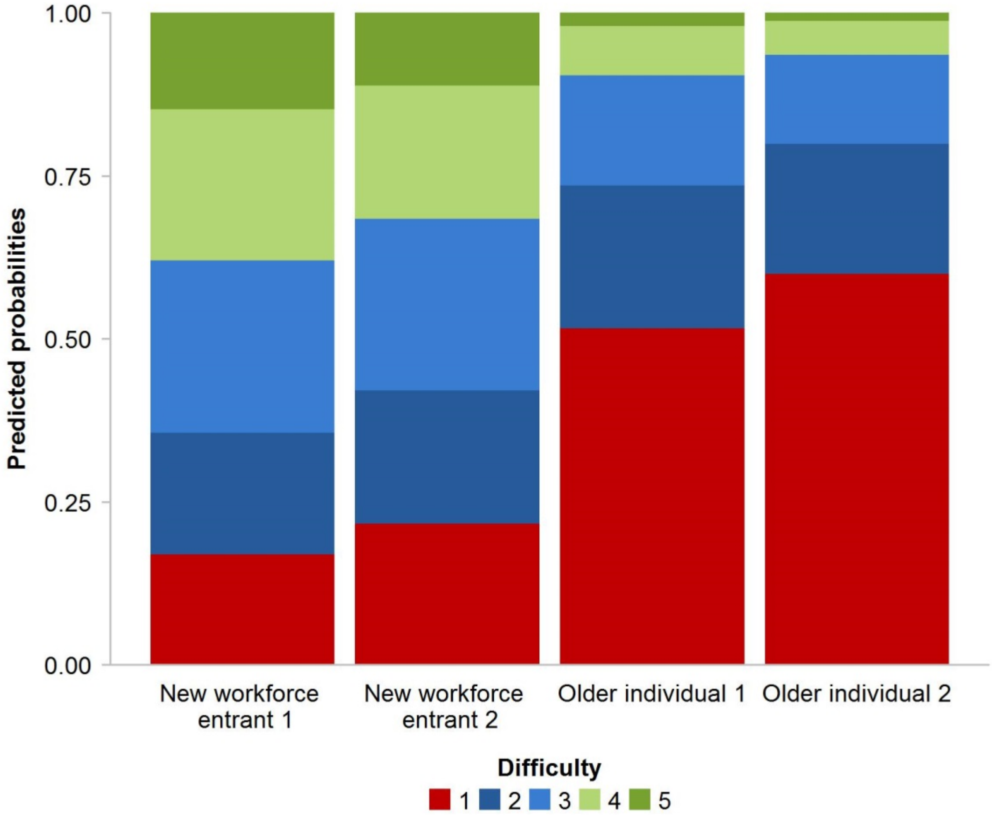



[a] 515 respondents chose existing funds and these observations have been removed. Three observations were exact duplicates and have been removed. Three respondents experienced technical problems in their surveys and have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem. Two respondents failed to submit a difficulty score.

Ordered probit regressions (with matching) were conducted to estimate treatment effects and disentangle the effect of observables on the difficulty score. After matching, this regression used 1069 responses. Annex A.10 provides the full details about this regression.

Figure 29 presents predictions for the same set of hypothetical persons of interest. Respondents with existing funds appeared on average to have found the task easier. The predicted proportion of respondents reporting a difficulty of one are much larger for respondents who already have a fund than for new workforce entrants, whereas the proportion of those reporting five are much smaller than for the new workforce entrants. This might be due to respondents simply opting for their existing fund or a learning effect. This effect is both economically and statistically significant.

Figure 29 **Predicted probabilities of reporting a given difficulty**
Hypothetical persons



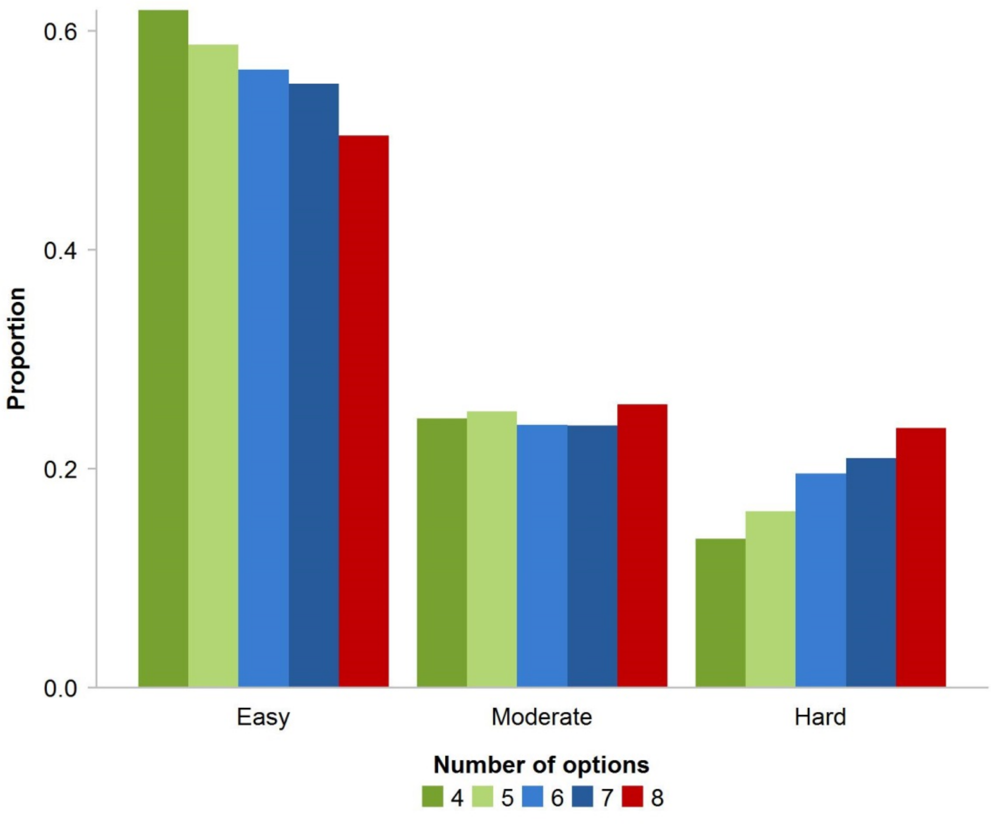

Higher levels of education appeared to moderate the difficulty of the task; most of these coefficients were both economically and statistically significant. This seems reasonably intuitive; educated respondents may be more likely to have an understanding of the task, but they may also put more effort in and experience more difficulty.

Higher financial literacy appeared to be associated with lower difficulty scores, although most of these coefficients were not statistically significant. Household income brackets appeared to be relatively unimportant economically and not statistically significant, which contrasts with the correlation of household income with nomination rates.

## Comparing the different shortlists by difficulty

Figure 30 shows the proportion of respondents in each sub-group choosing each category of difficulty. To improve clarity, responses have been aggregated into three difficulty groups instead of five. Selections of existing funds have been removed to preclude the option of choosing the existing option without much thought.

Figure 30 **Difficulty of choice experiment by number of options[a,b,c]**
N = 1120

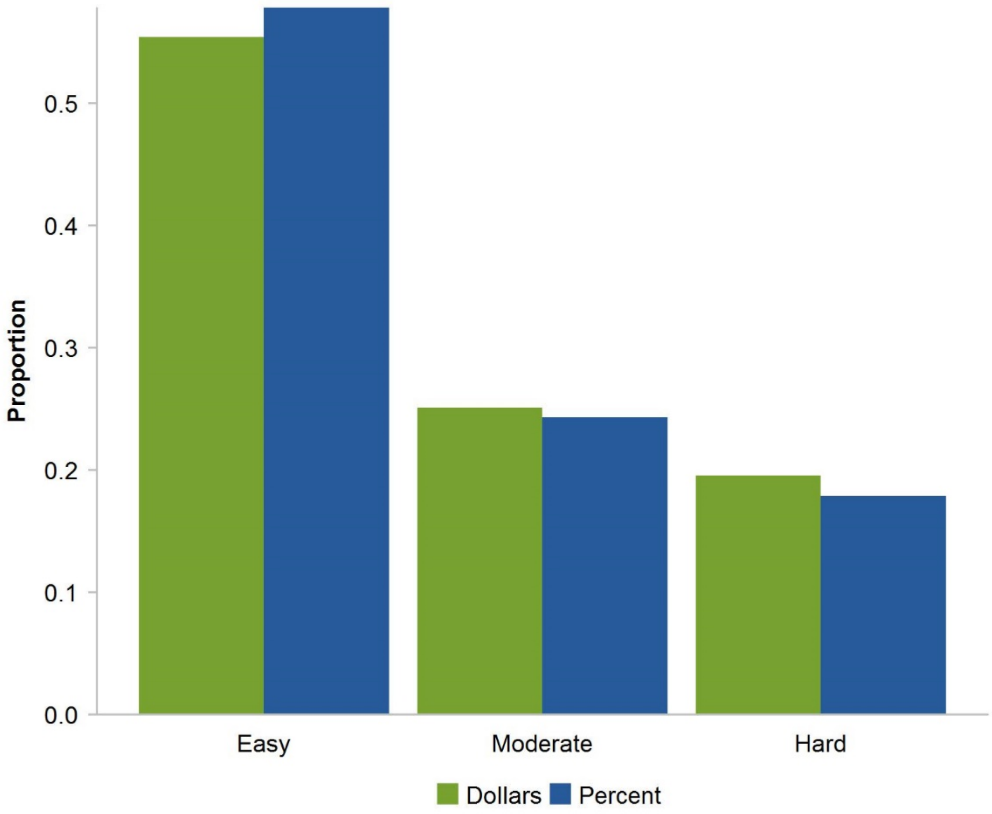



**a** 515 respondents chose existing funds and these observations have been removed. **b** Difficulties of one and two have been grouped into easy, a difficulty of three is assigned to moderate, and difficulties of four and five are assigned to hard. **c** Three observations were exact duplicates and have been removed. Three respondents experienced technical problems in their surveys and have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem. Two respondents failed to submit a difficulty score.

Having a greater number of options appeared to increase the proportion of respondents that found the task hard, although the size of the effects is relatively small. Similarly, figure 31 shows the proportions of respondents reporting each of the three aggregated difficulty categories broken down by the style of presentation of information on investment returns and fees. In this case, there appeared to be even less variation between the groups.

Figure 31 **Difficulty of choice experiment by dollar/per cent[a,b,c]**

N = 1120

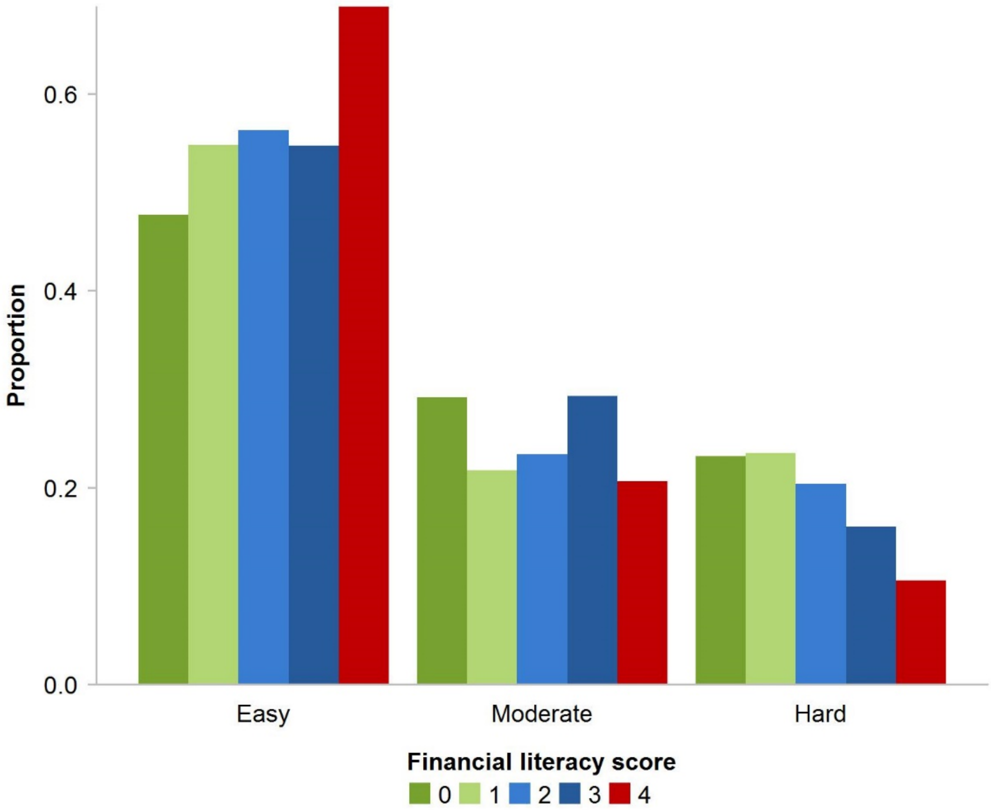



**a** 515 respondents chose existing funds and these observations have been removed. **b** Difficulties of one and two have been grouped into easy, a difficulty of three is assigned to moderate, and difficulties of four and five are assigned to hard. **c** Three observations were exact duplicates and have been removed. Three respondents experienced technical problems in their surveys and have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem. Two respondents failed to submit a difficulty score.

Last, figure 32 illustrates a significant positive effect on the difficulty of the task for those respondents with higher levels of financial literacy.

Figure 32 **Difficulty of choice experiment by financial literacy score[a,b,c]**

N = 1096




[a] 515 respondents chose existing funds and these observations have been removed.[b] Difficulties of one and two have been grouped into easy, a difficulty of three is assigned to moderate, and difficulties of four and five are assigned to hard. [c] Three observations were exact duplicates and have been removed. Three respondents experienced technical problems in their surveys and have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem. Two respondents failed to submit a difficulty score. 28 respondents did not answer all financial literacy questions and were thus also removed.

An ordered probit regression with matched data was conducted to determine if the method of presentation — dollar or percentage — had any influence on difficulty. Consistent with the figures there was neither an economically nor statistically significant effect. Annex A.11 provides the full details about this regression.

An ordered probit regression on raw data was conducted to determine if the number of options presented had any influence on difficulty. 1426 observations were used for this analysis. Accounting for observables, the regression showed that overall the number of options presented in the shortlist had minimal impacts on the difficulty of the task. None of the treatment indicators are economically or statistically significant. Annex A.11 provides the full details about this regression.

# 7 What did participants choose and what influenced those choices?

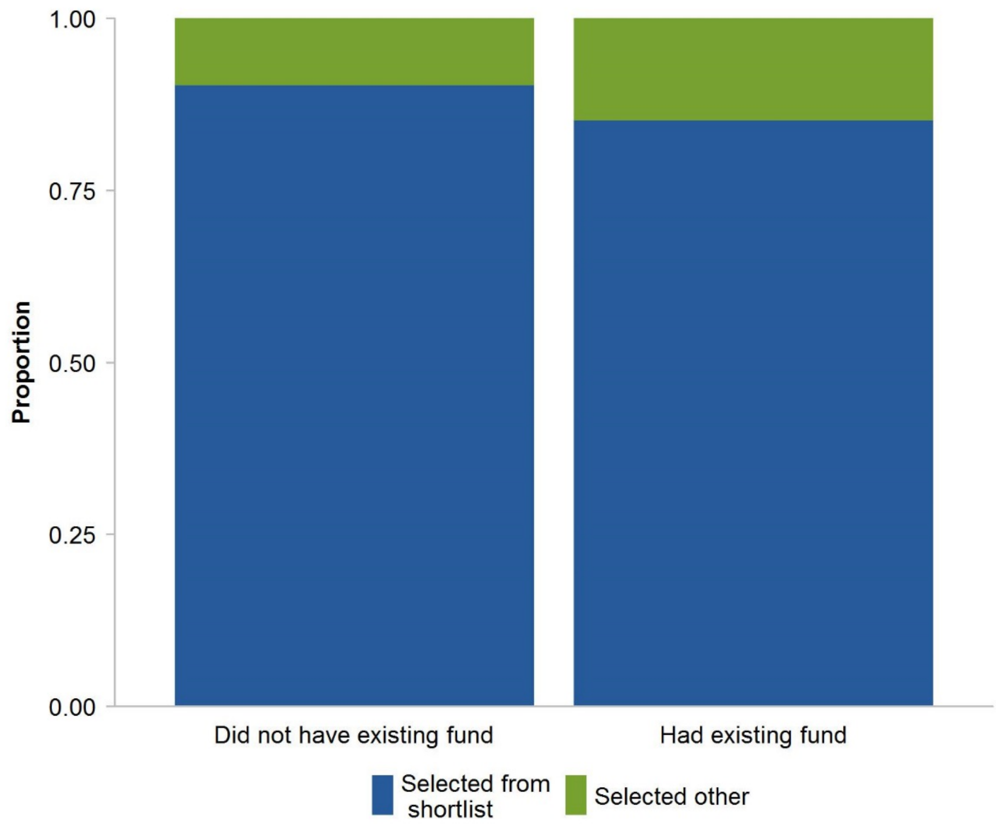
Respondents in the treatment groups were presented with a series of 4–8 products they could choose. Accompanying each product were the fund brand, and a randomly selected block of fund characteristics including past returns, fees, risk and target of the product. These features facilitate the exploration of factors influencing assisted choice selections. This section focuses on the assisted choice group, because the Commission does not have enough usable data regarding factors that may have influenced unassisted choice selections.

Because respondents were also allowed the option of nominating any fund or product outside of the list, it is first important to examine whether there are any self-selecting mechanisms behind respondents nominating from the shortlist or outside of it. For example, a potential factor influencing respondents to choose outside of the shortlist might be whether they have an existing fund or not.

The proportion of those who selected a product not on the shortlist is higher in the group with existing funds than without (figure 33). Despite this, the proportion of respondents with an existing fund choosing a product not on the shortlist is still relatively low. On the other hand, there is a small proportion of individuals without an existing fund who choose outside of the shortlist.

Figure 33 **Proportions of selections in the treatment group[a,b]**

N = 1509




**a** The data used for this analysis did not include nomination skippers, since the actual nominations were required. **b** Three observations were exact duplicates and have been removed. Three respondents experienced technical problems in their surveys and have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem.

Figure 34 shows the fund selections of respondents (for those choosing from the shortlist) based on the rank of risk, fees and returns of the selected products.[15] Box 3 discusses the ranking procedure in further detail.

15 Selections based on targets were not plotted as only 5 per cent of respondents in this figure mentioned targets in their open-ended responses.

## Box 3 Ranking fund characteristics

It is important to recall that the fund characteristic sets were randomised. Since the treatment was also split into different groups based on the number of options respondents were presented, this implies that most respondents only saw a random selection of the full array of fund characteristic sets. For example, a respondent who was presented with four options would only see four of the possible fund characteristic sets, while only respondents presented with eight options saw all the possible fund characteristic sets.

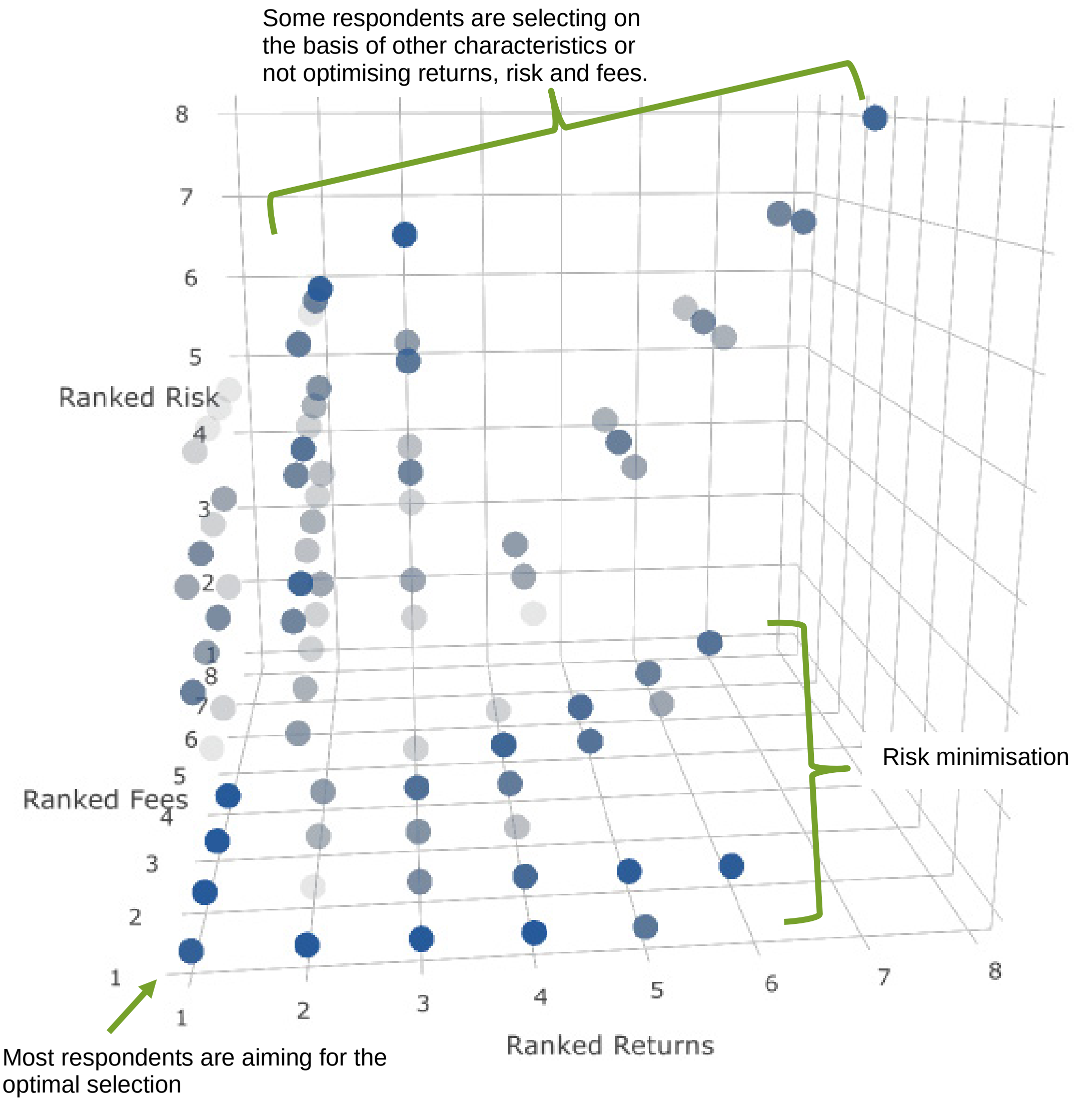
This means that it is not possible to directly compare fund selections based on the exact level of risk, fee or return, because the results may be biased. For example, many individuals may randomly see only the lowest return sets of fund characteristics, such that their selection might look like they intentionally chose a low return fund, but this might have been due to not having access to the higher return funds.

To address this issue, the Commission has ranked the risk, fees and returns for each individual's possible choices. The ranking is a sports-styled ranking, such that one is the best or highest rank. In the case of ties, both products receive the higher of the two possible ranks. High risk is assigned to lower ranks, high fees are assigned to lower ranks, and high returns are assigned to higher ranks.

Figure 34 shows a large clustering of observations around the corner with rank one on fees, risk and return. This suggests that many respondents are choosing funds on the basis of minimising fees and risk, and maximising returns. Second, considering the intensity of observations on the bottom plane of the figure, a large proportion of respondents are minimising risk. However, there is a substantial amount of heterogeneity in selections. There are a number of tails shooting away from the minimal fee and risk, and maximum return corner. This result remains true even if the sample is restricted to only those who receive the maximum financial literacy score and exhibit sufficient survey effort. This may suggest that other characteristics such as service quality or branding matter in addition to the presented characteristics, or that informed respondents are not making good choices. It is not possible to distinguish between (or directly test) these possibilities with the available data.

Figure 34 **Fund selections based on the rank of risk, fees and returns**[a,b]

N = 1287




[a] Three observations were exact duplicates and have been removed. Three respondents experienced technical problems in their surveys and have been removed. 205 respondents did not participate in the choice experiment. 272 respondents were removed due to the skipping nomination problem.[b] The Commission has used points with differing intensities in colour. The greater the intensity the more respondents chose the corresponding option.

## Exploring the effect of fund level effects on decision making

Figure 34 shows a large amount of heterogeneity in product choices. A key aspect that was missing in that analysis is consideration of fund level unobserved effects. These are factors such as member services, insurance and brand recognition that are different for each fund brand and unobserved from the Commission's point of view. These factors might have driven respondent choices in addition to the performance characteristics presented to respondents in the experiment. This section evaluates the role of those effects. Table 4 presents two blocks of fund characteristics and the ranks for each of the characteristics. As before, lower fees, higher returns, lower risk and higher targets are associated with higher ranks.

In table 4, fund characteristic set five has at least equal or superior characteristics compared with fund characteristic set one. Thus, in theory, if respondents choose fund characteristic set one over fund characteristic five, then there are only two plausible explanations. The first is that the respondent made a mistake, and would have chosen a different product had they properly evaluated and been provided with another chance. The second is that there are unobserved fund-level effects that drive the selection. Although analysis cannot identify when a mistaken choice has been made, it also seems unreasonable to expect all such respondents to systematically make mistakes. The effect of unobserved fund-level effects can be identified by including an indicator variable for the brand of the fund chosen in a regression to test if that had an influence on decisions. Taken together, this strategy could identify the role of fund-level unobservable effects in choices.

Table 4 **Fund characteristic set ranks[a]**

| *Fund characteristic set* | *Rank of returns* | *Rank of targets* | *Rank of risk* | *Rank of fees* |
|---|---|---|---|---|
| 1 | 7 | 2 | 1 | 8 |
| 5 | 5 | 2 | 1 | 6 |

**a** In the case of ties, the highest of the possible ranks are assigned. For example if three fund characteristic sets have the equal highest returns, all are assigned with a rank of first, and then the fund with the fourth highest returns is assigned a rank of fourth. **b** Six other blocks of fund characteristics are not relevant to the discussion at hand, and have been omitted.

A probit regression was conducted to estimate the fund-level unobservable effects. An indicator for someone choosing fund characteristic set one in the presence of fund characteristic set five was regressed on indicators for each of the eight funds, a survey effort measure, an indicator for whether the respondent had an existing fund and demographic variables. 730 observations were used for this regression. A log likelihood ratio test showed that this model was significant. Annex A.12 provides the full details about this regression.

Figure 35 shows the rates of selecting fund characteristic set one in the presence fund characteristic set five, for the hypothetical persons of interest (table 1). The figure shows there is some variation in these rates across different types of respondents. Thus some of the heterogeneity observed in product selection in figure 34 can be explained by variation in demographics.

Figure 35 **Heterogeneity across respondents**

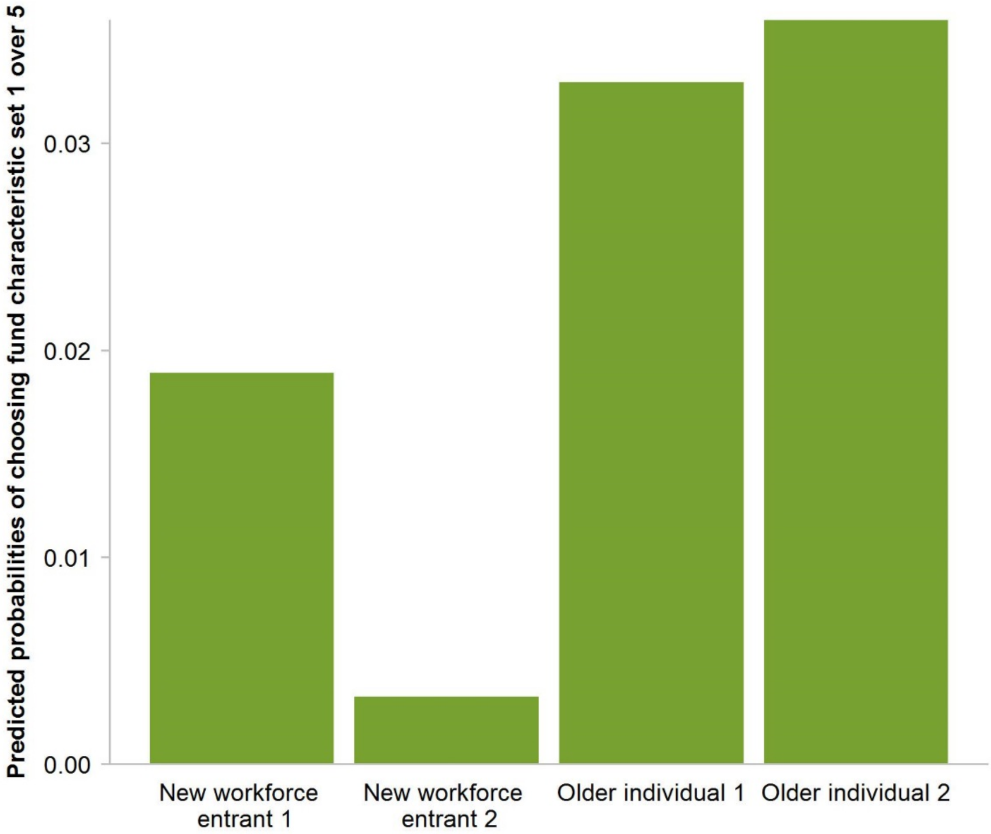



Figure 36 shows the rate of selecting fund characteristic set one in the presence of fund characteristic set five by the fund brand chosen. The predictions are based on older individual 2. The figure shows that there can be heterogeneity between funds, however the fact that these probabilities scale in accordance with the bars presented in figure 35 suggest that fund brands have a lesser role in explaining the heterogeneity in product selection observed in figure 34.

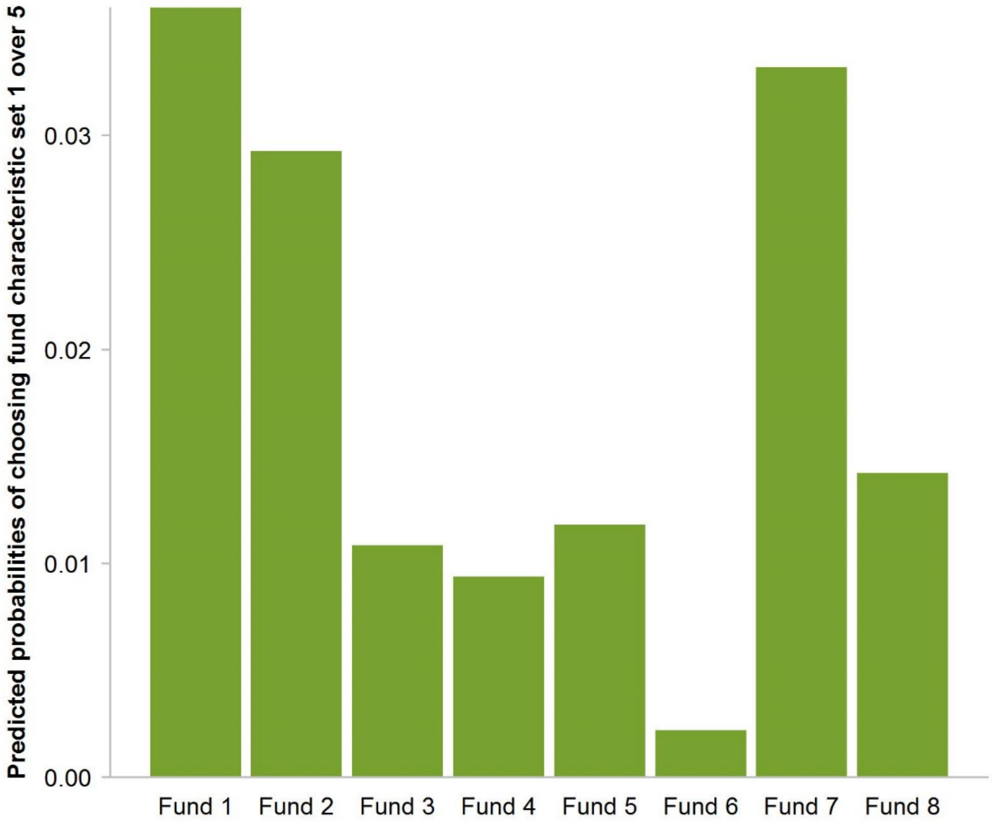
Regarding model coefficients more specifically, most of the fund level unobservable effects are economically significant, while only one is statistically significant. Considering the sample sizes, this is to be expected.

Having an existing fund increases the predicted probability and is economically significant, but not statistically significant. This is consistent with respondents choosing existing funds as a reason to ignore the fund characteristics. Respondents in older age brackets have economically significant effects of increasing the rates, with the 25–34 age category also being statistically significant.

Income bracket appeared to be relatively unimportant, while level of education is an important determinant. Having education of at least a year 12 standard reduces the rates and these effects are economically and statistically significant. Financial literacy also reduces these rates. These effects tend to be smaller compared to having better education, though still economically and statistically significant.

Figure 36 **Heterogeneity across funds[a]**

For older individual 2




[a] Plotting the chart for a different respondent mostly has a scaling effect. For example if the figure were based on new workforce entrant 1, there would be the same heterogeneity, but the overall magnitudes of effect would be smaller corresponding with the smaller predicted probabilities.

# TECHNICAL ANNEXES

# A Technical annexes for the survey

## A.1 Regression conventions

This section outlines the notational conventions used throughout the annexes.

### Regression conventions

- Respondents are indexed by $i = 1, \dots, N$.
- Unless otherwise specified, $D$ is a matrix of demographics with associated coefficient vector $\beta_d$. The matrix includes the age bracket, household income bracket, gender (an indicator, where a value of 1 represents a female), level of highest education, the binary survey effort measure for respondents (and where a value of 1 represents sufficient survey effort) and whether respondents have a current fund or not.
- $f_i$ is a vector of financial literacy score dummies associated with scores of 1 to 4 for respondent $i$ and the associated vector of coefficients is $\beta_f$. This means that these estimates are relative to a respondent scoring 0. Occasionally, $f_i$ will be the analogous numerical variable, with coefficient $\beta_f$. Unless otherwise specified, $f_i$ will be a categorical variable.
- $T_i$ is a treatment indicator. For example, whether someone has been randomly allocated to the assisted employee choice group and $\beta_T$ is the associated coefficient. $\beta_{T,f}$ is used to represent the vector of interaction coefficients between financial literacy score and the treatment effect.
- Unless otherwise specified, $\alpha$ represents the constant (intercept).
- Unless otherwise specified, $\varepsilon_i$ will be used to represent errors.
- $\Phi$ is used to represent the standard normal cumulative distribution function.
- Reported p-values for coefficients come from Wald tests.

Further, references to a ‘section X’ refer to the relevant document in the main text (About the survey and the results). And references to a ‘section A.X’ refer to internal sections within this document.

## A.2 List experiment

This section provides an explanation of the list experiment. In particular, it details how they work, how they are analysed, and where they have been used (section 1). Full details about the results are also provided (section 4).

## A.2.1 List experiments, and the analysis of

A 'list experiment' is a technique that is commonly used by psychologists and political scientists to gauge attitudes where respondents may not feel comfortable providing honest answers to questions asked directly.[16] List experiments are easily understood as a special randomised control trial (RCT). In the control group a list of 'control' statements are presented and respondents are asked to state how many of those statements they agree with. The treatment group is presented with a similar task, but with the list consisting of both all the control statements and an additional 'sensitive' statement. The sensitive statement is the statement for which estimates about the proportion agreeing with are sought. An example of a list experiment group conducted in our survey is presented in figure A.1.

After conducting a list experiment, the number of statements that a respondent agrees with is available. The proportion of respondents agreeing with the sensitive statement can then be identified by comparing the average number of statements agreed with in the control group against the treatment group. The control group average provides the baseline (reference), thus the difference between the control and treatment group must be attributed to the population proportion of respondents who agree with the sensitive statement. The Commission used linear regression consistent with the approach used throughout to estimate these differences and their standard errors. The regression equations were of the form:

$$Number\ of\ statements\ agreed\ with = \alpha + \beta_T T + \varepsilon$$

Note that $\alpha$ can be interpreted as the average number of statements that a respondent agrees with in the control group.

16 Some examples include Sniderman, Tetlock and Piazza (1992) where racial prejudice was measured and in Holbrook and Krosnick (2010) where voter turnout at presidential elections in the US was estimated based on survey methods.

Figure A.1 **Example of list experiment**


### A.2.2 Assumptions of list experiments and associated tests

It should be noted that along with the traditional assumptions imposed by regression and RCTs, there are two additional assumptions imposed in the usage of a list experiment. These are:

- *No 'design effects'*. This means that a participant's responses to the control items should not change in the presence of the sensitive item
- *Honest responses*. This means that participants do not lie about the sensitive item.

The first assumption can be tested using methods proposed by Blair and Imai (2012). The test results are presented in table A.1, with discussion in section A.5.3 relevant for understanding the table. Overall there does not appear to be evidence suggesting there are design effects.

Table A.1 **List experiment design effect test results**

| *Sensitive Statement* | *p-value using raw data* | *p-value using matched data* |
|---|---|---|
| I chose a fund at random | 1.0000 | 0.9453 |
| I felt uneasy thinking about my retirement | 1.0000 | 0.7787 |
| I was annoyed at how much time and energy it took to choose a fund | 1.0000 | 0.3056 |
| I really gave no thought to how my choices might affect my account balance | 1.0000 | 0.3223 |
| I chose a fund that sounded familiar to me | 1.0000 | 0.9522 |
| I trusted other people to make the decision for me | 0.9595 | 0.5437 |
| I wished my employer had just recommended a fund suitable for my lifestage and circumstances | 1.0000 | 1.0000 |
| I was influenced by funds' advertising | 1.0000 | 1.0000 |
| I didn't fully understand how the super contributions I made now would affect my retirement income later | 1.0000 | 0.1565 |
| I didn't really know what a "super fund" was | 1.0000 | 0.5266 |
| I felt overwhelmed by the number of choices before me | 1.0000 | 0.2876 |
| I chose the first fund I came across | 0.7054 | 1.0000 |
| I didn't end up making a fund choice at all | 1.0000 | 1.0000 |
| I couldn't care less about super | 1.0000 | 1.0000 |
| I didn't really know what it meant to say a fund had a certain level of "risk" | 1.0000 | 0.4019 |
| I was actually annoyed at having to choose a fund | 0.6334 | 0.9837 |
| I didn't really know what sorts of information I should consider in making my choice | 1.0000 | 1.0000 |
| I pretended to care about the decision, but I didn't really care at all | 1.0000 | 1.0000 |
| I felt overwhelmed by the amount of information I was supposed to consider in making my choice | 1.0000 | 1.0000 |
| I went with a super fund I already had, without considering any other information | 0.8967 | 0.7397 |
| I didn't fully understand who was supposed to make super "contributions", or how | 0.7499 | 1.0000 |
| I had a good understanding of what a fund's "asset allocation" referred to | 0.9610 | 0.5332 |
| I didn't understand the meaning of super "returns" and how they were calculated | 1.0000 | 1.0000 |
| I didn't really know what super was for | 0.1780 | 1.0000 |

The second assumption is important, but as Blair and Imai (2012) suggest, hard to test for directly. Their solution is to instead directly model the possibility of lies. This modelling would require significant additional research. Considering the anonymisation and relatively uncontentious nature of the sensitive statements, the Commission considered that it seems unlikely that there would be systematic widespread lying, and will assume this assumption is satisfied.

### A.2.3 List experiment results

This section presents two set of results: the ‘raw’ results estimated from the full sample, and a ‘corrected’ set of results estimated from a smaller, more balanced sample. The large number of statements being tested meant the sample allocated to each branch was significantly smaller than the sample at large. This led to some of the raw estimates being produced from an ‘unbalanced’ — with respect to demographic characteristics — sample. A matching technique (detailed in section A.5.2) was used to re-balance the samples of each branch and produce the corrected set of estimates.

While the raw estimates may come from a somewhat unrepresentative sample, the re-balancing process involved reducing the number of observations, and therefore reduced the ‘power’ of the estimates. The risk of the raw estimate is that the estimates will be biased due to the imbalance and therefore be incorrect for our sample of respondents.[17] The risk of an underpowered estimate is that if the experiment were re-ran with randomly re-drawn samples, there is a higher chance of the result being a false positive, and also a higher risk of a true positive with an inaccurate magnitude being produced.[18] Put differently, although the matched set of results may not be replicable due to relatively low power, they are still accurate representations based on our sample of respondents, whereas the raw set of results may not even be correct about our sample of respondents. The Commission is most confident about results that are relatively invariant between using matched or raw data.

As can be observed, there are substantial differences in both statistical and economic significance across the two sets of estimates (table A.2). With regards to statistical significance, 22 of the 24 statements see a reduction (that is, an increased p-value). The fact that the re-balancing saw much fewer statistically significant results implies that the concern of false positives from an underpowered estimate is likely to be minor in this case.

With regards to economic significance, the average change in magnitude was a slight increase of 0.63 per cent. This reflects the relatively even distribution of changes around zero.

[17] See Rubin (2008) and Ho et al. (2007) for discussions about the importance of minimising imbalance.

[18] See Button et al. (2013) for a discussion about why underpowered studies are more likely to produce false positives, and Gelman and Carlin (2014) for a discussion about why underpowered studies may produce incorrect estimates.

## Table A.2 List experiment regression results

Raw and corrected samples

| | *Raw sample* | | | | *Corrected Sample* | | | |
|---|---|---|---|---|---|---|---|---|
| Sensitive statement | Per cent agree | Std. error | p-Value | n | Per cent agree | Std. error | p-Value | n |
| **Selection engagement and understanding** | | | | | | | | |
| I trusted other people to make the decision for me | 52.2 | 10.3 | 0.00 | 301 | 52.0 | 17.5 | 0.00 | 122 |
| I was influenced by funds' advertising | 35.1 | 10.9 | 0.00 | 301 | 46.7 | 16.7 | 0.01 | 122 |
| I felt overwhelmed by the number of choices before me | 35.3 | 11.6 | 0.00 | 293 | 36.7 | 19.5 | 0.06 | 107 |
| I didn't end up making a fund choice at all | 44.7 | 11.5 | 0.00 | 298 | 36.7 | 18.1 | 0.05 | 120 |
| I was actually annoyed at having to choose a fund | 14.2 | 12.4 | 0.25 | 298 | 36.7 | 19.6 | 0.06 | 120 |
| I felt overwhelmed by the amount of information I was supposed to consider in making my choice | 30.5 | 12.1 | 0.01 | 289 | 32.7 | 19.4 | 0.10 | 105 |
| I wished my employer had just recommended a fund suitable for my lifestage and circumstances | 32.9 | 12.2 | 0.01 | 301 | 28.3 | 18.9 | 0.14 | 122 |
| I went with a super fund I already had, without considering any other information | 25.2 | 11.7 | 0.03 | 289 | 28.2 | 17.7 | 0.11 | 105 |
| I was annoyed at how much time and energy it took to choose a fund | 35.2 | 12.0 | 0.00 | 292 | 25.5 | 18.4 | 0.17 | 120 |
| I chose a fund that sounded familiar to me | 48.6 | 12.7 | 0.00 | 301 | 24.6 | 20.3 | 0.23 | 122 |
| I chose the first fund I came across | 8.5 | 11.3 | 0.45 | 293 | 22.5 | 20.2 | 0.27 | 107 |
| I chose a fund at random | 36.0 | 11.9 | 0.00 | 292 | 15.1 | 17.3 | 0.39 | 120 |
| **System engagement and understanding** | | | | | | | | |
| I didn't fully understand who was supposed to make super "contributions", or how | 50.2 | 11.1 | 0.00 | 288 | 56.0 | 20.8 | 0.01 | 81 |
| I didn't fully understand how the super contributions I made now would affect my retirement income later | 43.1 | 11.8 | 0.00 | 293 | 54.5 | 21.0 | 0.01 | 107 |
| I had a good understanding of what a fund's "asset allocation" referred to | 31.9 | 11.8 | 0.01 | 288 | 47.1 | 20.3 | 0.02 | 81 |
| I didn't really know what it meant to say a fund had a certain level of "risk" | 26.9 | 10.1 | 0.01 | 298 | 41.7 | 17.3 | 0.02 | 120 |
| I didn't really know what super was for | 49.9 | 12.1 | 0.00 | 288 | 41.4 | 21.7 | 0.06 | 81 |
| I was actually annoyed at having to choose a fund | 14.2 | 12.4 | 0.25 | 298 | 36.7 | 19.6 | 0.06 | 120 |
| I really gave no thought to how my choices might affect my account balance | 44.8 | 11.0 | 0.00 | 292 | 35.4 | 16.6 | 0.03 | 120 |
| I didn't really know what a "super fund" was | 22.7 | 11.5 | 0.05 | 293 | 24.6 | 20.3 | 0.23 | 107 |
| I didn't understand the meaning of super "returns" and how they were calculated | 48.5 | 12.3 | 0.00 | 288 | 24.4 | 24.0 | 0.31 | 81 |
| I couldn't care less about super | 23.9 | 10.5 | 0.02 | 298 | 21.1 | 16.4 | 0.20 | 120 |
| I felt uneasy thinking about my retirement | 22.7 | 11.6 | 0.05 | 292 | 15.1 | 17.5 | 0.39 | 120 |

## A.3 Relevance of financial literacy scores

This section documents the regression analysis conducted to support the construction of the financial literacy measure (section 2). The criterion that is being tested is if the financial literacy measure seems to be a reasonable measurement. Regressing household income on demographics, including the financial literacy score, allowed the Commission to gauge whether this criterion is likely to be satisfied. Intuitively respondents with higher financial literacy should have a higher household income.

The regression equation is:

$$Income\ bracket\ midpoint = \alpha + \beta_f f + \beta_d D + \varepsilon$$

Where:

- The household income bracket midpoints are computed as outlined in section 2
- $D$ includes in this case the age bracket, level of highest education and gender of respondents.

Note that regressions are run for the cases where $f$ is treated as a numerical variable (assuming linearity) (table A.3), and as a categorical variable (relaxing linearity) (table A.4).

As these regressions are only used for establishing correlations; goodness of fit and model significance have not been considered. Note that the estimate of $19 000 quoted in the main text of the effect of an additional point of financial literacy score was computed as the average effect from the financial literacy score estimates in the regression with financial literacy as a categorical variable.

Table A.3 **Household income ($'000) and financial literacy[a,b,c,d,e]**

Financial literacy as a numerical variable

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 4.3118 | 0.4380 | *** |
| **Financial literacy score** | 0.7961 | 0.1005 | *** |
| **Age bracket** | | | |
| 25-34 | 1.1521 | 0.4060 | *** |
| 35-44 | 2.2607 | 0.4039 | *** |
| 45-54 | 0.7816 | 0.4053 | ** |
| 55-64 | -0.3096 | 0.4223 | |
| **Female** | -0.3228 | 0.2478 | |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.8664 | 0.4334 | ** |
| Certificate/Diploma | 1.2783 | 0.4053 | *** |
| Bachelor | 2.7695 | 0.4246 | *** |
| Graduate Diploma | 3.802 | 0.6011 | *** |
| Postgraduate | 5.0783 | 0.5286 | *** |

**a** The constant reflects a male respondent in the 15-24 age bracket who scored 0 on the financial and super literacy questions, and has a less than year 12 education. **b** The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined. **c** Respondents with level of highest level of education categorised as 'other' based on their open responses have been removed. **d** 1992 observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively.

Table A.4 **Household income ($'000) and financial literacy[a,b,c,d,e]**

Financial literacy as a categorical variable

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 4.0123 | 0.4897 | *** |
| **Financial literacy score** | | | |
| 1 | 1.4356 | 0.4175 | *** |
| 2 | 1.812 | 0.4050 | *** |
| 3 | 2.6604 | 0.4144 | *** |
| 4 | 3.4417 | 0.4609 | *** |
| **Age bracket** | | | |
| 25-34 | 1.1573 | 0.4064 | *** |
| 35-44 | 2.2649 | 0.4043 | *** |
| 45-54 | 0.8099 | 0.4062 | ** |
| 55-64 | -0.292 | 0.4230 | |
| **Female** | -0.3137 | 0.2495 | |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.8632 | 0.4336 | ** |
| Certificate/Diploma | 1.2544 | 0.4057 | *** |
| Bachelor | 2.7677 | 0.4246 | *** |
| Graduate Diploma | 3.7833 | 0.6018 | *** |
| Postgraduate | 5.087 | 0.5296 | *** |

[a] The constant reflects a male respondent in the 15-24 age bracket who scored 0 on the financial and super literacy questions, and has a less than year 12 education. [b] The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined. [c] Respondents with level of highest level of education categorised as 'other' based on their open responses have been removed. [d] 1992 observations were used for this regression. [e] ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

## A.4 Validity of the survey effort measure

This section considers two points of evidence that support the measure of survey effort (section 2) as a valid measurement of the effort respondents exerted in their surveys.

### A.4.1 Does the survey effort measure correlate with the time taken during the choice experiment?

Although it could be argued that using the time taken during the choice experiment as a survey effort measure would be problematic, it is still useful as a sense check of the measure. Regardless of how noisy the time taken as a survey effort measure might be, it should still be the case that, on average, a respondent who takes longer to complete the survey should have spent more effort on the survey. To examine this statement a regression of the form:

$$Time\ taken\ during\ choice\ experiment = \alpha + Survey\ effort\ measure + \varepsilon$$

was conducted, where the survey effort measure referred to, takes the values of $0, 1$ and $2$ (table A.5).

As these regressions are only used for establishing correlations, goodness of fit and model significance have not been considered.

Table A.5 **Time taken during choice experiment (sec.) and survey effort[a,b,c,d]**

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 37.2685 | 20.4741 | * |
| **Survey effort measure** | | | |
| 1 | 19.3805 | 22.5412 | |
| 2 | 59.4015 | 21.5301 | *** |

**a** The constant reflects a respondent who did not answer any open ended response questions in the choice experiment. **b** Note that one respondent who took over 80 000 seconds to complete this section was removed from this regression. As the next longest respondent took just over 5000 seconds, this is a clear outlier, and the inclusion of the respondent would distort regression results. **c** 2123 observations were used for this regression.**d** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively..

Regression analysis showed that there is indeed such a correlation. Longer survey completion times are associated with more open ended questions in the choice experiment being answered. For example, someone who answered both open-ended choice experiment questions spent almost a minute longer completing the choice experiment on average. This suggests that survey effort is reflected in the number of open responses to some extent.

## A.4.2 Is the survey effort measure useful or spurious?

Another way to check for the validity of the measure is to check if the measure is likely to be spurious. To this end, a regression was conducted to determine if the survey effort measure was correlated with other demographics. The regression specification was:

$$P(Survey\ effort\ sufficient = 1) = \Phi\left(\alpha + \beta_f f + \beta_d D\right)$$

For this probit regression, the binary survey effort measure has been used. For this measure, Respondents who answered at least one open-ended response question in the choice experiment are assigned a value of 1 and 0 otherwise. This is a probit regression with the dependent variable being whether someone exerted a moderate survey effort. This is because survey effort is best thought of as a categorical variable, but an ordered regression would not be readily interpretable for this context. As these regressions are only used for establishing correlations, goodness of fit and model significance have not been considered.

The regression showed that there is a lack of an association between the survey effort measure and income brackets and most level of highest education categories (table A.6). There appeared to be some explanatory value in the survey effort measure. That is, the survey effort measure variation was not fully explained by other predictors.

Taken together these sense checks showed that our survey effort measure is useful. It balances between minimising the exclusion of participants while remaining relevant to measuring survey effort.

Table A.6 **Survey effort correlates**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---:|---:|---:|
| **Constant** | -0.2239 | 0.1350 | * |
| **Age bracket** | | | |
| 25-34 | -0.2661 | 0.1013 | *** |
| 35-44 | -0.3634 | 0.1013 | *** |
| 45-54 | -0.4083 | 0.1010 | *** |
| 55-64 | -0.2357 | 0.1060 | *** |
| **Female** | 0.4113 | 0.0615 | *** |
| **Household income bracket ($ p.a)** | | | |
| 20 000 – 39 999 | 0.1717 | 0.1270 | |
| 40 000 – 59 999 | 0.013 | 0.1234 | |
| 60 000 – 79 999 | 0.0881 | 0.1242 | |
| 80 000 – 99 999 | 0.0018 | 0.1286 | |
| 100 000 – 129 999 | -0.0185 | 0.1308 | |
| 130 000 – 199 999 | 0.1428 | 0.1369 | |
| > 200 000 | -0.1078 | 0.1674 | |
| **Financial literacy score** | | | |
| 1 | 0.2359 | 0.0987 | *** |
| 2 | 0.5692 | 0.0978 | *** |
| 3 | 0.827 | 0.1025 | *** |
| 4 | 0.9623 | 0.1158 | *** |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.0042 | 0.1048 | |
| Certificate/Diploma | 0.1617 | 0.0991 | |
| Bachelor | -0.0129 | 0.1039 | |
| Graduate Diploma | 0.3471 | 0.1543 | ** |
| Postgraduate | 0.1434 | 0.1320 | |

**a** The constant reflects a male respondent in the 15-24 age bracket who scored 0 on the financial and super literacy questions, and has a less than year 12 education. **b** The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined. **c** Respondents with level of highest level of education categorised as 'other' based on their open responses have been removed. **d** 1998 observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

## A.5 Suitability of data — balance, power and representativeness

This section provides a full discussion of some of the issues outlined in section 2. To contextualise this section, core concepts of causal inference are reviewed. Techniques and figures for addressing each of these concepts are then presented.

### A.5.1 First principles of valid causal inference

Causal inference asks whether X causes Y. Policy design often relies on knowing about causation between two variables. Although the questions are conceptually simple and seem easy to study, valid causal inference is notoriously hard.

To motivate this understanding, consider a policy example posed by Angrist and Pischke (2009). The research question is do hospitals make people healthier? Heuristically this should be true (although it is possible increased public exposure to disease could make the reverse true), but in addition to confirming this effect, having an understanding of the quantitative impact would allow policy makers to assess if relatively more spending should go to hospitals or preventative health programs for example. The most intuitive method of studying this question would be to compare the health outcomes of those who have been to a hospital recently to those who have not. Angrist and Pischke drew on a national American health outcomes survey to show that under this intuitive analysis, the conclusion would be that hospitals substantially reduce health outcomes. As they highlight, the key problem with this analysis is that healthy people may have self-selected into the group that did not recently go to a hospital, while sick people may have self-selected into the group that did recently go to a hospital. In effect, the analysis compares the health outcomes of healthy people who have not been to a hospital against those who are sick and have been to the hospital. In other words this analysis is subject to selection bias.

Moving away from hypotheticals, Angrist and Pischke also highlight the evolution of the labour economics literature on the effect of government-subsidised training programs on earning outcomes. Early studies ignored potential selection bias issues arising from workers of low earnings potential self-selecting into the training programs and showed that these training programs had a negative impact on the earnings of workers. More recent studies that utilised randomised experiments confirmed that the earlier studies were subject to selection bias and that these training programs typically support earning potentials.

The latter example hints at a common characteristic of valid causal inference. The use of a randomised experiment, or more popularly, an RCT. The 'randomisation' in a randomised experiment refers to the random assignment of participants between control (a comparison group) and treatment groups (the group that an intervention or policy change is applied to). In the labour economics case training program applicants were randomly allocated into a control group that did not receive the training program, and a treatment group that underwent the training program. The random allocation in these cases solves the selection bias problem as self-selection is not allowed.

It is important to stress however that randomisation is not necessary nor sufficient for ensuring a valid causal inference. The ultimate goal is that the analysis take place on an apples-to-apples comparison basis – this is often referred to as having *balanced* control and treatment groups. In each of the examples above, the main issue was that groups of people with clearly different characteristics were being compared. Randomisation is applied to bring each of the groups as close as possible to perfectly *balanced*. This does not mean that each individual in the groups must look the same, it just means that the 'average' person (averaging over observable characteristics) in each of the treatment and control groups looks as similar as possible. Many studies are based on 'natural experiments' that are valid because sufficient balance has already been achieved. On the other hand, application of randomisation may not always achieve sufficient balance. This may be due to the uneven sizes of the control and treatment groups for example.

Balance is not the only key to a valid causal inference. Consider an experiment with three participants in a treatment group and three participants in a control group. How credible would an analyst regard the experiment? Valid causal inference requires that the sample sizes in the analysis are large enough such that the analysis can conclude the results are not due to chance – this is often referred to as an adequately powered analysis. Underpowered analysis are problematic because not only is it less likely that the analysis will be able to detect a treatment effect, but for similar reasons, a significant finding is more likely to be a false positive.[19] In many fields, a lack of power has led to a lack of replicability in results (Button et al. 2013; Ortmann nd; Russ and Gelman nd).

Finally, a valid causal inference is also dependent on the context of the analysis. Consider an experiment that features only workers close to retirement age, but is conducted for the purpose of evaluating the effects of training programs for young unskilled workers. Even though such an experiment might have been executed properly and might correctly identify the treatment effect of training programs for old workers, the results would not be credible for the target population because the samples are not 'representative'.

## A.5.2 Balance and matching

In practice, perfect balance in any analysis is unlikely. There may have been an aspect of randomisation that was not accounted for, or the analysis may have been based on observational data as parts of the Commission's survey was. Yet even small imperfections

[19] This argument is specified in more detail in Button et al. (2013). Underpowered studies mean that the rate of true positives is low. But the rate of false positives is fixed (holding the level of significance and type of test constant). Thus conditional on seeing a significant result, it is more likely that the result is a false positive. Recent research shows that there is also a greater risk of wrong estimates (Gelman and Carlin 2014).

in balance can lead to relatively large bias. Therefore, attempts should be made toward getting the balance as close to ideal as possible (Ho et al. 2007). This section explores the methods adopted by the Commission to improve balance, reduce selection bias and hence improve the credibility of analysis where possible.

In many econometric applications, parametric and modelling-based adjustments are made to reduce selection bias, some of these techniques include the usage of instrumental variables, Heckman corrections, or outright structural modelling. The Commission has decided against using model-based adjustments because applying model-based adjustments also requires a proper development of underlying theory, which seems difficult to conceptualise in the types of questions and experiments being conducted.

For this reason, the Commission has employed non-parametric methods of reducing selection bias. These methods are commonly referred to as matching methods. To understand why, consider the following process. Starting with the original treatment and control samples, look for the two most similar observations or respondents (drawing one from control and one from treatment) and put the two observations into a new pool of matched data. This process is repeated until there are no more similar observations. Analysis is conducted with the new pool of matched data.

All matching methods involve a ‘matching step’ at some stage, although how that is carried out may look quite different. Nonetheless, all of them have the same goal — to reduce selection bias by ensuring that similar groups are being compared. An additional advantage of matching methods is that they reduce the dependence of analysis on the modelling techniques chosen. The reason is because matching methods reduce the ‘lack of overlap’ problem. Suppose that a treatment observation is added to a dataset. The observation is just unique enough that they do not dramatically alter the balance of the treatment and control groups, but is different enough such that there are no similar subjects in the control group. Any modelling technique, regardless of how sophisticated will then be forced to extrapolate when using this information for the estimation of the treatment effect (and in doing so affect predictions that do not require extrapolation). Matching ensures that such an observation would be pruned away so that models do not have to extrapolate. Box A.1 summarises some of the popular matching methods.

## Box A.1 Matching Methods

**Nearest Neighbour Matching**

Nearest neighbour methods use a scalar distance metric to compare observations. A popular distance metric is the Mahalanobis distance. Heuristically, for each observable, this technique measures the distance between the observations in terms of their standard deviations accounting for the correlation structure in the sample. The next step involves using multivariate Pythagorean identities to aggregate the distances on different dimensions into a scalar distance. Regardless of the distance metric, observations are 'matched' when the distance is 'small' enough.

**Propensity Score Matching**

Propensity score matching is more commonly employed in observational studies or partial randomisation. The first step in propensity score matching is to compute the probability a given observation will be in the treatment group — a propensity score, most commonly using a logistic regression. Matching is then conducted on the basis of the propensity scores and observations are 'matched' when the distance between propensity scores is 'small' enough.

**Coarsened Exact Matching (CEM)**

CEM starts with first coarsening the dataset. Variables that are continuous are discretised into an arbitrary number of categories decided by the analyst. Already discrete variables can be coarsened further, by grouping and reducing the number of categories. Observations can then be categorised into 'strata' by the exact combination of their coarsened values.

Observations are then 'exactly' matched. Unlike the former methods, because all variables in the coarsened dataset are discrete, matching can be done by finding a control and treatment unit in the same strata, and there is no need for a scalar distance metric.

**Advantages of CEM**

Some literature appeared to support CEM as superior to other matching methods (Iacus, King and Porro 2011). Relative to other matching methods, CEM has a number of advantages. From an analysis point of view, there can be a trade-off to consider when using matching methods. On the one hand improving balance is important, on the other hand, reducing observations can lead to larger standard errors and lower power. By controlling the amount of coarsening, the analyst can control how much they wish to improve balance and how many observations they want to retain. In contrast other matching methods maximise balance, sometimes at the expense of low sample sizes because the analyst has little control over the algorithm.

Another advantage of CEM is that the matching step is exact and eschews the need for an abstract scalar distance metric. A major problem with abstract measures is that counter-intuitive matches could be possible. For example, an observation thatlooks similar to an observation with only small differences on a number of observables like age, income and financial literacy might be superseded by another observation with notably more education, but few differences otherwise as the abstract measure is agnostic about many real world nuances.

*Sources*: King et. al. (2011); Iacus et. al. (2011).

The Commission used the coarsened exact matching (CEM) method. Propensity score matching does not really make sense in this survey as the subjects did not have the opportunity to self-select. Nearest neighbour matching is difficult to control and dropped too many observations when tested.

The coarsening resulted in many variables having fewer categories, while still maintaining sufficient contextual relevance (figure A.2). Coarsening was not done for the age bracket variable, as the age brackets of 10 years are already quite coarse.

Figure A.2 **Matching variables and the coarsening step[a]**

| Variables to match on | Original groupings | Coarsened groupings |
| --- | --- | --- |
| Female | No change | No change |
| Age bracket | No change | No change |
| Income bracket | < 20 000<br>20 000 – 39 999<br>40 000 – 59 999<br>60 000 – 79 999<br>80 000 – 99 999<br>100 000 – 129 999<br>130 000 – 199 999<br>> 200 000 | < 40 000<br>40 000 – 79 999<br>80 000 – 129 999<br>> 130 000 |
| Financial literacy | 0<br>1<br>2<br>3<br>4 | 0 - 1<br>2<br>3 - 4 |
| Level of highest education | Less than year 12<br>Year 12 or equiv<br>Cert/Dip<br>Bachelor<br>Graduate Dip<br>Postgrad. | Less than year 12<br>Year 12 or equiv &<br>Cert/Dip<br>Bachelor<br>Graduate Dip & Postgrad |

**a** In the case of the choice experiment groups, matching also included an indicator for whether someone currently had a fund. This was not included in the list experiment as doing so would've further reduced the already small sample sizes. Moreover the indicator was not relevant to the analysis techniques following the matching.

A measure for pre- and post-imbalance matching is required to examine how much the matching procedure helps. The measure is called a $L_1$ distance and measures imbalance in the joint distributions of the control and treatment groups, by measuring the percentage of the distributions that do not overlap. Note that it is a measure that can only be used to assess improvements as it is a relative measure. This means it does not make sense to compare $L_1$ distances across different datasets. $L_1 = 1$ corresponds to the distributions being completely separate, while $L_1 = 0$ represents complete overlap of the distributions. A $L_1$ of 0.6 means that 40% of the densities overlap, thus lower values of $L_1$ are better. The $L_1$ measure can also be used to compare univariate distributions. The technical details of the measure are left to box A.2.

### Box A.2 **Measuring imbalance**

The $L_1$ distance is a measure of differences between the multivariate empirical distributions for the treatment and control groups. This means that it is a measurement that looks across the whole picture including each observable and distributional differences as opposed to just differences in means.

The first step to constructing this measurement is considering the bin sizes of the histogram that will be generated. As per Iacus, King and Porro, the Commission adopts the default stance of choosing the median bin width based on all possible bin widths in the raw data.

Next define $H(X_k)$ as the set of distinct values generated by binning on variable $k = 1, \dots, K$. Then the multidimensional histogram constructed from these bins and data is the set of cells generated by the Cartesian product $H(X_1) \times H(X_2) \times \dots \times H(X_K) \coloneqq H(\boldsymbol{X})$.

Now let $f$ and $g$ be the empirical probability mass functions for the treatment and control groups and $f_{l_1 \dots l_k}$ and $g_{l_1 \dots l_k}$ are the probability masses corresponding to particular cells of the empirical probability mass function, indexed by the possible bins.

The multivariate imbalance measure is then:

$$L_1(f, g) = \frac{1}{2} \sum_{l_1 \dots l_k \in H(X)} \left| f_{l_1 \dots l_k} - g_{l_1 \dots l_k} \right|$$

In words, this means to sum up all the absolute differences in empirical probability masses for a multidimensional histogram bin and then divide by two.

By holding fixed the discretisation process, the measure can also be computed after matching for comparison purposes.

*Source*: Iacus, King and Porro (2011).

The steps to the Commission's approach in assessing and improving balance can thus be summarised as:

1. Apply CEM to improve balance of observables. The observables that were balanced on included age bracket, gender, level of highest education, financial literacy and household income bracket. This implicitly excludes observations from analysis that did not have all of these measures.
2. Assess the improvement in balance using $L_1$ measures. This process was applied to each of the list experiment branches, and some analyses for the choice experiment (table A.7).

Table A.7 **Matching outcomes**

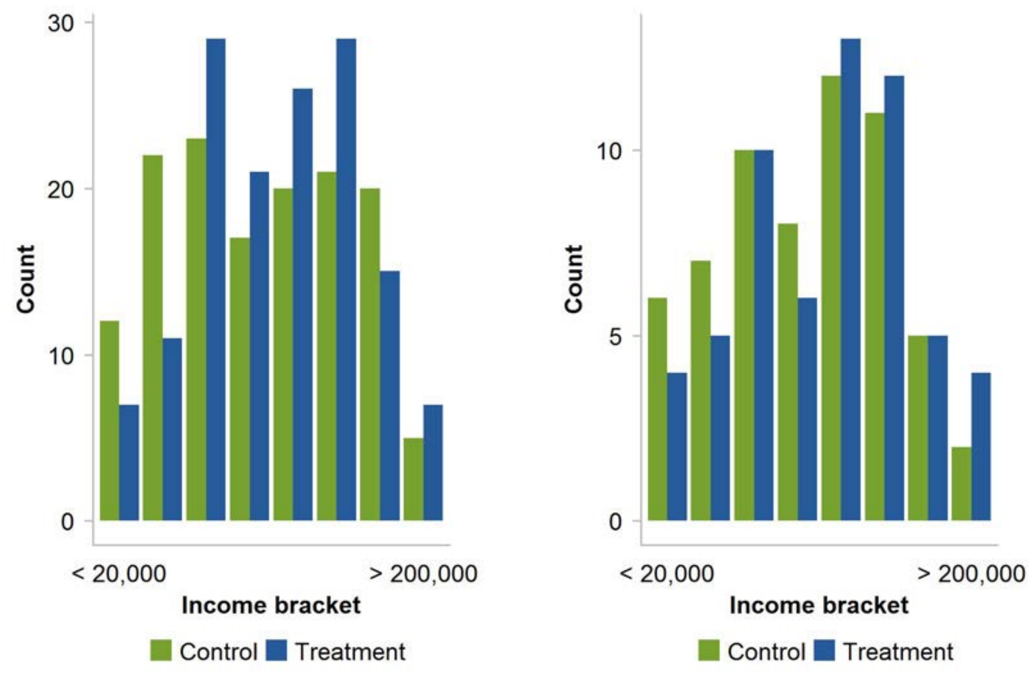
| *Subject* | *Pre-matching $L_1$* | *Post-matching $L_1$* | *Subject* | *Pre-matching $L_1$* | *Post-matching $L_1$* |
|---|---|---|---|---|---|
| **List experiment groups A and B** | | | | | |
| Overall | 0.8748 | 0.7240 | Income bracket | 0.1635 | 0.1148 |
| Female | 0.0357 | 0.0000 | Financial literacy score | 0.1562 | 0.0410 |
| Age bracket | 0.0421 | 0.0000 | Level of highest education | 0.1030 | 0.0710 |
| **List experiment groups C and D** | | | | | |
| Overall | 0.8739 | 0.6635 | Income bracket | 0.1853 | 0.2503 |
| Female | 0.1309 | 0.0000 | Financial literacy score | 0.1324 | 0.0873 |
| Age bracket | 0.0960 | 0.0000 | Level of highest education | 0.1331 | 0.0519 |
| **List experiment groups E and F** | | | | | |
| Overall | 0.9201 | 0.8173 | Income bracket | 0.0960 | 0.1538 |
| Female | 0.1013 | 0.0000 | Financial literacy score | 0.0860 | 0.1346 |
| Age bracket | 0.0840 | 0.0000 | Level of highest education | 0.1558 | 0.1010 |
| **List experiment groups G and H** | | | | | |
| Overall | 0.8919 | 0.7500 | Income bracket | 0.1594 | 0.1806 |
| Female | 0.0827 | 0.0000 | Financial literacy score | 0.0563 | 0.0583 |
| Age bracket | 0.0925 | 0.0000 | Level of highest education | 0.1418 | 0.0889 |
| **List experiment groups I and J** | | | | | |
| Overall | 0.0504 | 0.0000 | Income bracket | 0.1083 | 0.0673 |
| Female | 0.0365 | 0.0000 | Financial literacy score | 0.0772 | 0.0351 |
| Age bracket | 0.1506 | 0.1901 | Level of highest education | 0.0504 | 0.0000 |
| **List experiment groups K and L** | | | | | |
| Overall | 0.1088 | 0.0000 | Income bracket | 0.0540 | 0.0542 |
| Female | 0.0808 | 0.0000 | Financial literacy score | 0.1082 | 0.1958 |
| Age bracket | 0.1181 | 0.2375 | Level of highest education | 0.1088 | 0.0000 |
| **Choice experiment; assisted and unassisted employee choice** | | | | | |
| Overall | 0.8026 | 0.6107 | Income bracket | 0.0522 | 0.0330 |
| Female | 0.0645 | 0.0000 | Financial literacy score | 0.0507 | 0.0481 |
| Age bracket | 0.0004 | 0.0000 | Level of highest education | 0.0736 | 0.0356 |
| **Choice experiment; per cent and dollar** | | | | | |
| Overall | 0.7219 | 0.5737 | Income bracket | 0.0296 | 0.0084 |
| Female | 0.0594 | 0.0000 | Financial literacy score | 0.0503 | 0.1010 |
| Age bracket | 0.0061 | 0.0000 | Level of highest education | 0.0513 | 0.0000 |

It is worth noting that the Commission did not conduct matching for all analyses. In particular, matching was not conducted for analyses involving the breakdown of treatment by number of options presented. This is because matching for each of the five groups would result in samples too small, as each of the treatment groups would require being balanced to each other. On the other hand, balancing was conducted in the list experiment despite relatively small samples because the imbalance issues appeared tangible (table A.7).

Figure A.3 highlights the tangibility of balance and the improvements that can be made from applying CEM. The charts show the number of respondents within each income bracket between control and treatment groups for a particular branch in the list experiment. The left hand side chart represents the original data, while the right represents the data after applying CEM. It can be seen that the control group has a systematically larger group of low household income individuals compared with the treatment group, which features more higher household income individuals. After applying CEM it can be seen that the distributions much more closely align. Heuristically, if the raw data were used we might expect to underestimate the proportion of respondents who agree with a statement like ‘I chose a fund at random’ because the treatment group (where the average individual has a higher income) might tend to disagree with the statement.

Figure A.3 **Potential improvements from balancing**

Pre (LHS) post (RHS) applying CEM



## A.5.3 Sample sizes and power simulations

Unlike with balance, the Commission's approach to power was more binary. That is, are we adequately powered? It should be noted upfront that computing power is always an arbitrary task. Many assumptions need to be made to conduct an analysis, so computing power is intended to be a rough sense-check rather than a precise test.

To investigate the power in analyses, the Commission has conducted power simulations based on regression techniques commonly used throughout the appendix, namely linear and probit regressions. To fix things, power simulations have been conducted to cover each of the experiments. In the choice experiment case, probit regressions are conducted, regarding the effect of the assisted employee choice model on nomination rates of funds. In the list experiment case, linear regressions are conducted in estimating a hypothetical proportion of respondents agreeing with the sensitive statement.

Simulations start by using informed guesses of parameters (such as the coefficient estimate and standard deviation of the dependent variable), and simulating the analyses on fake data to compute the power — the proportion of times the analysis returns a statistically significant result.

### Choice experiment simulations

The pseudo-code for the choice experiment simulations proceeds as follows:

1. Set the simulation parameters. The treatment and constant parmeters were informed by corresponding estimates on the actual data. The regression used in this case was of the form:

$$P(Nominates) = \Phi(\alpha + \beta_T T_i)$$

Table A.8 presents the results of the regression. Table A.9 presents the power simulation parameters and results.

2. Randomly draw standard normal errors. Probit regressions are used, which normalise the parameters of the normal distribution to a standard normal distribution.
3. Generate the set of fake data to run the regression on. This includes generating the binary outcome variable, which is done so using the latent variable formulation of the probit regression and the treatment indicator. The latent variable formulation can be expressed as:

$$Y = \begin{cases} 1, Y^* > 0 \\ 0, otherwise \end{cases}$$

Where $Y^*$ refers to the latent variable corresponding to the nomination decision:

$$Y^* = \alpha + \beta_T T_i + \varepsilon_i$$

It is important to note that these coefficients are not interpreted as probabilities. For example, $\alpha$ does not represent the baseline probability.

To see this, recall the equivalence of the latent variable formulation and the generalised linear model formulation:

$$P(Y = 1) = P(Y^* > 0) = P(\alpha + \beta_T T_i < -\varepsilon_i) = P(\varepsilon_i < \alpha + \beta_T T_i) = \Phi(\alpha + \beta_T T_i)$$

Care is taken when setting coefficient values for the simulations.

4. Run a probit regression of the form:

$$\Pr(Nominates) = \Phi(\alpha + \beta_T T_i)$$

And save the p-value associated with the treatment coefficient.

5. Repeat for the number of iterations chosen. Then compute the power as the proportion of times the p-value is less than 0.1.

Table A.8 **Regression for simulation parameter selection**[a,b,c]
Choice Experiment

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 1.5682 | 0.1179 | *** |
| **Treatment** | 0.4247 | 0.1535 | *** |

[a] The constant reflects a control respondent.[b] 1069 observations were used for this regression, CEM matching was run for this regression.[c] ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively..

Under reasonably conservative parameter assumptions the power in choice experiment analyses is likely to be 'adequate' (table A.9).[20]

Table A.9 **Power simulation parameters and results**[a,b]
Choice experiment

| **Parameters** | | | |
|---|---|---|---|
| Number of Simulations | 5000 | | |
| Number of control subjects | 291[a] | Number of treatment subjects | 778[a] |
| Treatment effect size | 0.4247[b] | Constant coefficient | 1.5682[b] |
| ***Power*** | 0.8496 | | |

[a] These sample sizes come from the CEM – matched data used for the regression presented in table A.2. This is a conservative choice, because some other analyses use more observations. [b] These estimates come from the regression presented in table A.8. These estimates are conservative because control variables which reduce noise and increase power have not been accounted for.

20 In academia, the benchmark for adequate power is typically 80 per cent. That is, analyses have a 80 per cent chance of detecting a true effect when there is one (Gelman and Carlin 2014).

## List experiment simulations

The pseudo-code for the list experiment simulations are similar and proceeds as follows:

1. Set the simulation parameters (table A.10). The treatment, constant and standard deviation of dependent variable parameters were informed by corresponding estimates on the actual data (tables 2 and 3 in section 4).
2. Randomly draw normal errors with mean zero and standard deviation calibrated from the standard deviation observed in responses.
3. Generate the set of fake data to run the regression on. This includes generating the outcome variable that is the number of statements a respondent agreed with and the treatment indicator.
4. Run a linear regression of the form:

$$Number\ of\ statements\ agreed_i = \alpha + \beta_T T_i + \varepsilon_i$$

And save the p-value associated with the treatment coefficient.

5. Rinse and repeat for the number of iterations chosen. Then compute the power as the proportion of times the p-value is less than 0.1.

Table A.10 shows that generally speaking that many of the list experiment results are not adequately powered at the 80 per cent level. This underscores the Commission's conservative approach to analysis.

As noted earlier, low power leads to an increase in the rate of false positives. Many results lose significance or have increased p-values moving from the raw list experiment results to the matched list experiment results (table A.2), which might suggest that the concern of false positives is not salient.

However Gelman and Carlin (2014) suggest that when normally distributed, estimates are likely to be exaggerated when power is less than 50 per cent, and of the wrong sign when power is less than 10 per cent. Overall this suggests that list experiment results with relatively small effect sizes of about 10 to 20 per cent and large standard errors should not be taken very seriously. Other list experiment results are likely to be indicative, but may be exaggerated.

Table A.10 **Power simulation parameters and results**

List experiment

| **Number of Simulations** | 5000 | | |
|---|---|---|---|
| **Small samples, small effect size** | | | |
| Number of control subjects | 40[a] | Number of treatment subjects | 41[a] |
| Effect size | 0.15[b] | Constant coefficient | 1.0187[c] |
| Standard deviation | 1.0238 | | |
| ***Power*** | 0.1848 | | |
| | | | |
| **Small samples, average effect size** | | | |
| Number of control subjects | 40[a] | Number of treatment subjects | 41[a] |
| Effect size | 0.35[d] | Constant coefficient | 1.0187[c] |
| Standard deviation | 1.0238 | | |
| ***Power*** | 0.4694 | | |
| | | | |
| **Small samples, large effect size** | | | |
| Number of control subjects | 40[a] | Number of treatment subjects | 41[a] |
| Effect size | 0.50[e] | Constant coefficient | 1.0187[c] |
| Standard deviation | 1.0238 | | |
| ***Power*** | 0.721 | | |
| | | | |
| **Large samples, average effect size** | | | |
| Number of control subjects | 60[f] | Number of treatment subjects | 60[f] |
| Effect size | 0.35[d] | Constant coefficient | 1.0187[c] |
| Standard deviation | 1.0238 | | |
| ***Power*** | 0.5878 | | |

[a] These sample sizes come from the CEM matched data for branches K and L of the list experiment. This is a conservative choice, because other branches had more observations in both control and treatment. [b] This effect size corresponds to approximately the smallest effect sizes for our list experiments. [c] This is the average constant for all control branches. [d] This comes from the average of treatment effects for all list experiments. [e] This comes from the larger treatment effect estimates. A number of treatment effects are larger than 0.5 still. [f] These sample sizes come from the CEM matched data for branches G and H that had the most observations in both control and treatment groups across the branches. Note that with the exception of branches K and L, most other branches have similar, but slightly smaller sample sizes.

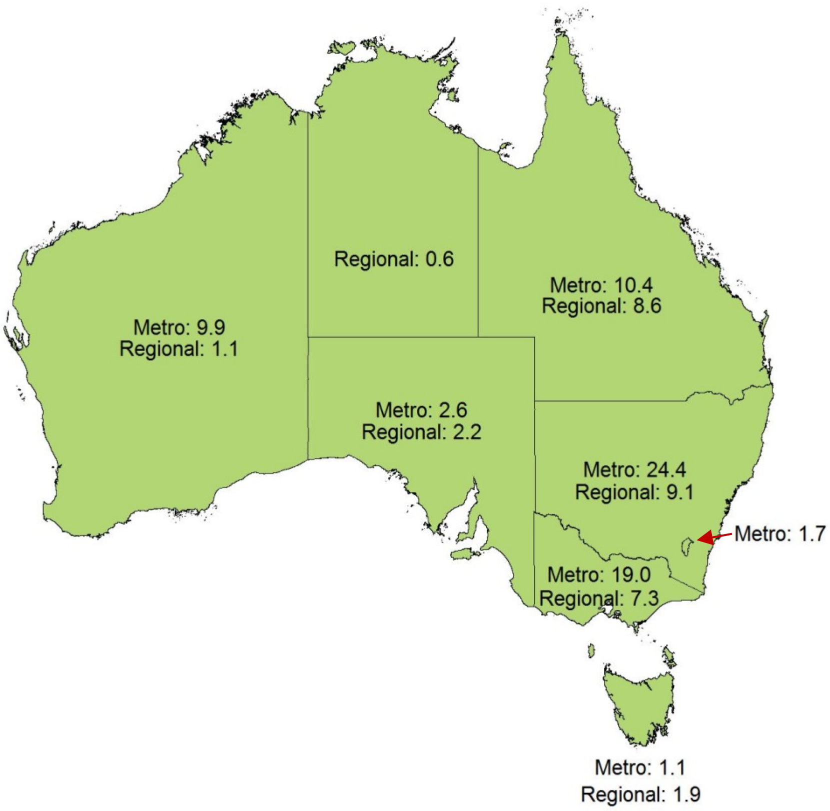
## A.5.4 Representativeness

The Commission's target population was primarily the wider Australian population. As demonstrated in section 2, the Commission is largely satisfied with the representativeness of the full sample. Applying CEM to the data should not adversely affect representativeness of the data since part of the aim of CEM is to reduce edge cases that are unlikely to have sufficient observations to enable appropriate estimation on.

Another aspect of representatives worth considering is that of younger Australians. Many policy relevant questions in this space pertain to the behaviour of younger Australians who are likely to be key stakeholders in policy reform regarding default products. To answer these questions a dataset representative of younger Australians is essential. In figures in this section, respondents who are between 15 and 24 are featured. Overall the spread of young Australians in our sample across regions seems representative (figure A.4). 52 per cent of 15-24 year olds in our sample are female.

Figure A.4 **15-24 Australians by region[a]**

N = 463

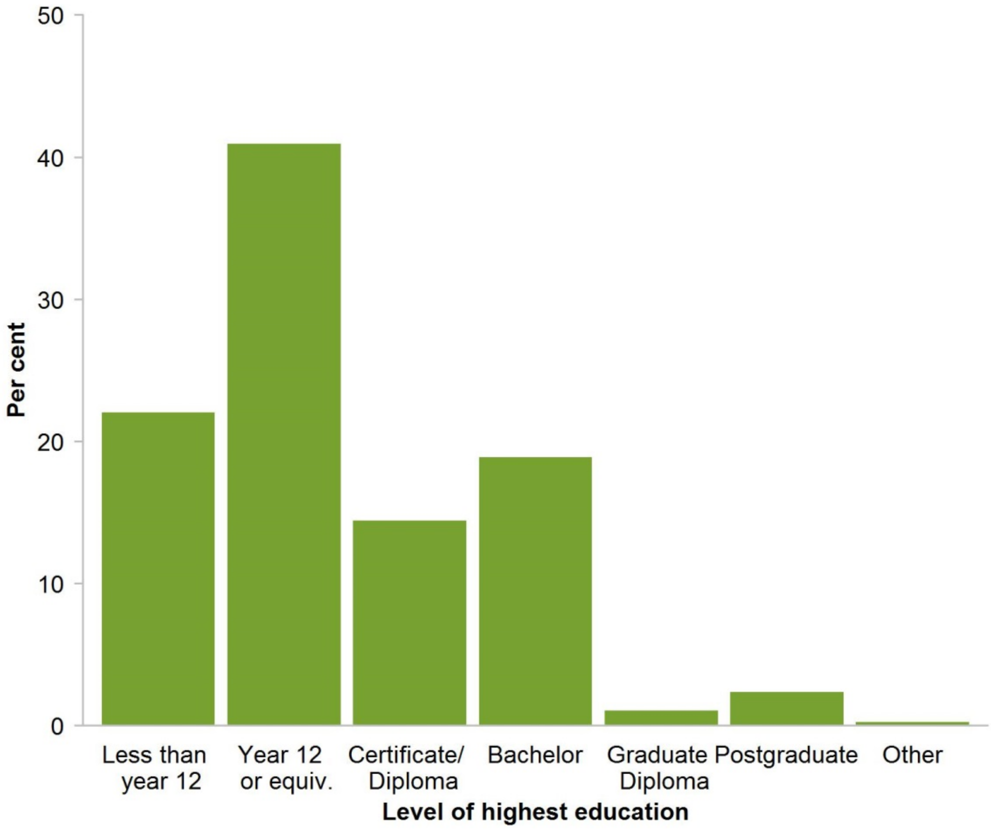



[a] One observation lacked the associated region. [b] Figures are reported as percentages, which may not sum to 100 due to rounding.

As expected, most young Australians in our sample have or are completing their schooling and a substantial number have higher qualifications (figure A.5).

Figure A.5 **15-24 Australians by level of highest education[a,b]**

N = 381

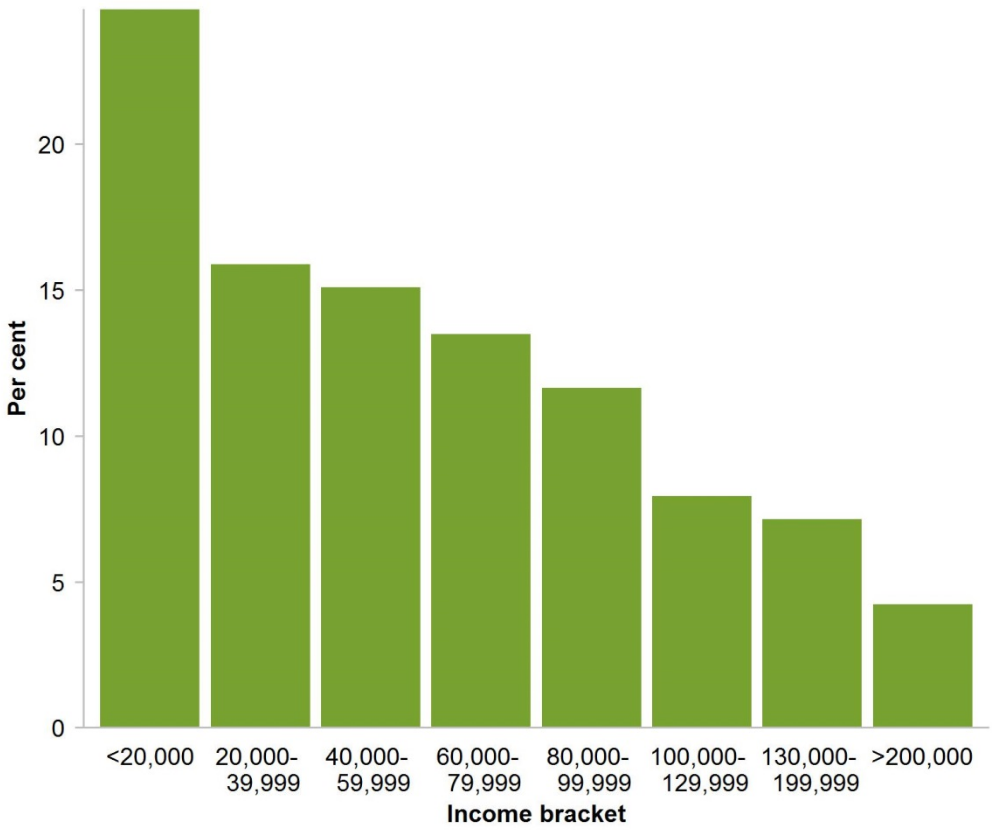



**a** ‘Less than year 12’ is the categories ‘primary school’, ‘some high school’ and ‘year 10 or equivalent’ combined. ‘Certificate/Diploma’ is ‘Trade certificate/Apprenticeship‘ and ‘Diploma/Associate degree’ combined. Postgraduate’ is ‘Masters’ and ‘PhD’ combined. **b** There are also 82 non-responses as this question was not mandatory.

Figure A.6 shows young Australians in the sample broken down by their household income bracket. A plausible issue with the income bracket variable is that individuals were asked about their household income as opposed to individual income. Therefore it could be possible that stay at home individuals may have larger household incomes than they otherwise would if they were by themselves. Although the figure does not provide any conclusive evidence towards a view, considering that the mean income appeared substantially lower than in figure 7 in section 2, this does not appear to be a substantial issue.

Figure A.6 **15 - 24 Australians by household income bracket**[a]

N = 378

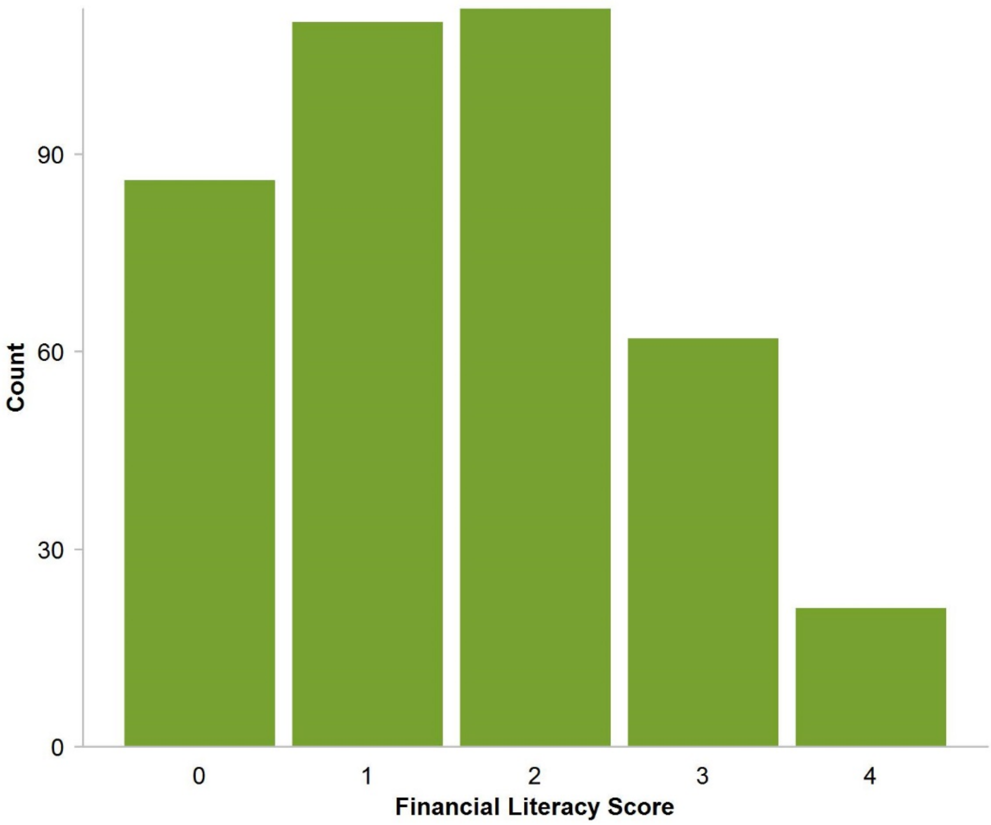



[a] There are also 85 non-responses for the associated question as this question was not mandatory.

As one might expect, the average financial literacy score for young Australians in our sample appeared lower than that in figure 8 of section 2 (figure A.7).

Figure A.7 **15-24 Australians by financial literacy**[a]

N = 391

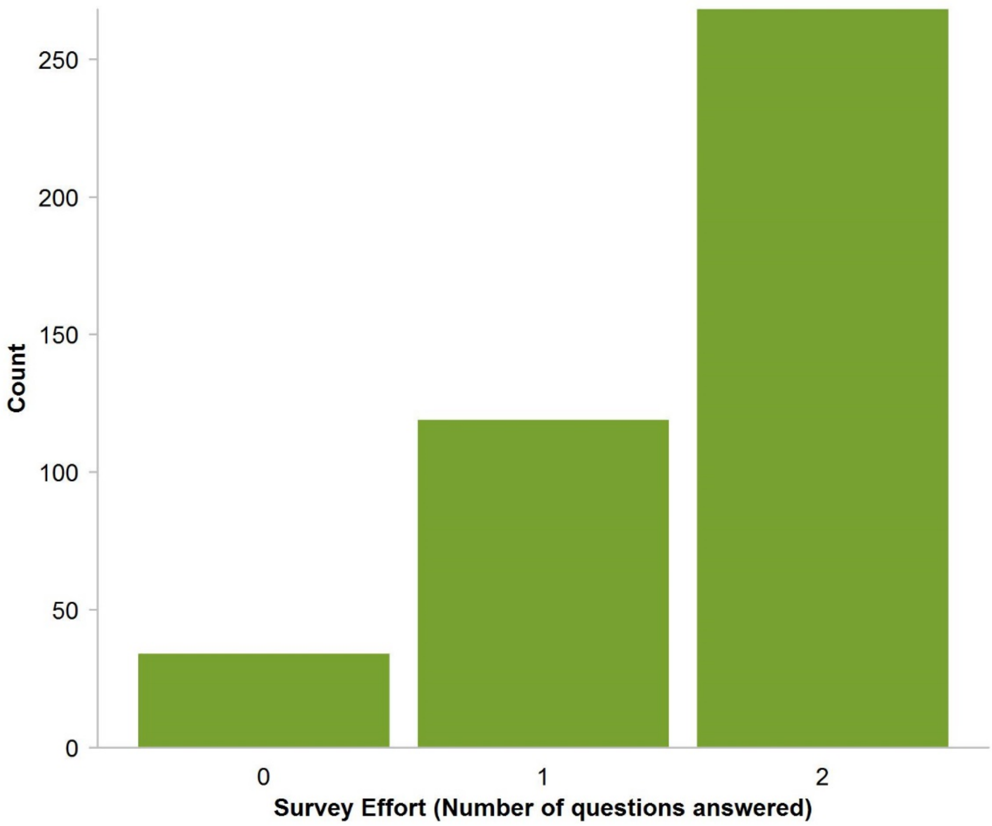



[a] There are also 72 respondents who did not answer all four financial literacy questions.

A potential concern might be that young Australians may exhibit systematically less effort due to lack of engagement or lack of understanding. Comparing figure A.8 and figure 9 in section 2, shows that this does not appear to be the case.

Figure A.8 **15-24 Australians by survey effort[a]**

N = 421




[a] 42 respondents dropped out before they participated in the choice experiment.

Taking the figures presented together, it appeared that the sample is reasonably representative of younger Australians. Thus with the appropriate techniques the Commission can report findings on the behaviour of younger Australians in the survey.

## A.6 Probit model for current defaulters

This section documents the regression detailed in section 3. The Commission estimated a probit model to explore the relationship between demographics and the propensity to default (table A.11). The regression equation is:

$$P(Nominates) = \Phi\left(\alpha + \beta_f f + \beta_d D\right)$$

Where in this case $D$ does not include the binary survey effort measure. This is because survey effort is not relevant from nominations in past experiences with nominations.

Table A.11 **Current defaulters probit model estimation results[a,b]**

Dependent variable = 1 if respondent defaulted and 0 otherwise.

| *Variable* | *Coefficient* | *Standard error* | *z-Score* | *p-Value* |
|---|---|---|---|---|
| Constant | 0.3159 | 0.1782 | 1.77 | 0.0760 |
| Female | 0.2045 | 0.0663 | 3.08 | 0.0020 |
| **Age bracket** | | | | |
| 25-34 | 0.1235 | 0.1161 | 1.06 | 0.2880 |
| 35-44 | 0.2382 | 0.1150 | 2.07 | 0.0380 |
| 45-54 | 0.1785 | 0.1156 | 1.54 | 0.1220 |
| 55-64 | 0.0364 | 0.1192 | 0.31 | 0.7600 |
| **Financial literacy score** | | | | |
| 1 | -0.0316 | 0.1308 | -0.24 | 0.8090 |
| 2 | 0.1015 | 0.1272 | 0.80 | 0.4250 |
| 3 | -0.0311 | 0.1277 | -0.24 | 0.8070 |
| 4 | -0.0115 | 0.1368 | -0.08 | 0.9330 |
| **Education** | | | | |
| Year 12 | 0.0137 | 0.1056 | 0.13 | 0.8970 |
| Certificate/diploma | -0.0780 | 0.1074 | -0.73 | 0.4680 |
| Bachelor degree | -0.0844 | 0.0936 | -0.90 | 0.3680 |
| Graduate cert/diploma | 0.0607 | 0.1448 | 0.42 | 0.6750 |
| Postgraduate | -0.3532 | 0.1234 | -2.86 | 0.0040 |
| **Household income ($'000 p.a)** | | | | |
| 20-40 | -0.1370 | 0.1644 | -0.83 | 0.4040 |
| 40-60 | -0.2463 | 0.1574 | -1.57 | 0.1180 |
| 60-80 | -0.4346 | 0.1577 | -2.76 | 0.0060 |
| 80-100 | -0.4319 | 0.1601 | -2.70 | 0.0070 |
| 100-130 | -0.1754 | 0.1631 | -1.08 | 0.2820 |
| 130-200 | -0.4277 | 0.1667 | -2.57 | 0.0100 |
| Over 200 | -0.6523 | 0.1975 | -3.30 | 0.0010 |
| **Other output** | | | | |
| Sample size | 1570 | Pseudo $R^2$ | 0.0284 | |

[a] Male was the base category for gender, 15-24 for age bracket, 0 for financial literacy score, 'less than year 12' for education, and less than $20 000 p.a. for household income. [b] The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined.

## A.7 Does the choice experiment influence close-ended responses?

In sections 3 and 4, footnotes discussed concerns about the sequencing of the choice experiment prior to the close-ended responses influencing participant's choices in those close-ended responses, such as for the product features that matter to them. This section explores these issues.

## A.7.1 Did the choice experiment influence nomination responses

Figure 13 in section 3 presented results on the method of fund nominations used by respondents from a close-ended question. One concern could be that respondents who have just undertaken an assisted nomination task, might be more inclined to remember (potentially incorrectly) having made an independent selection. To explore this concern, a multinomial logistic regression was conducted, with regression equation:

$$P(Nomination\,method = j) = \frac{e^{X\beta_j}}{1 + \sum_{k=1}^{K-1} e^{X\beta_k}}$$

Where $j$ is a set containing the responses:

- I made my own selection independent of anyone else
- I selected a fund recommended by someone else
- I used my employer’s default fund
- I used my existing fund
- other.

Blank responses are used as a pivot. The observables ($X$) used for this regression included age brackets, a gender indicator, income brackets, level of highest education and financial literacy scores. Survey effort was not included because there was insufficient data to estimate the full array of observables.

This regression allows us to examine if there are any changes in probabilities of choosing a particular nomination method depending on if a respondent was in the control or treatment group (table A.12). Only predictions have been shown as only effect sizes are relevant to this concern.

Overall from an economically significant point of view, there did not appear to be any major concerns. Most differences between control and treatment are well below 5 per cent. In all cases it appeared that fewer individuals are choosing ‘Other’ in the treatment group, which might suggest that the assisted nomination task may have helped respondents focus on the question a bit more than in the unassisted nomination task.

Table A.12 **Predicted probabilities of choosing different nomination methods**[a,b]

By control and treatment groups

| | *Control* | *Treatment* |
|---|---|---|
| **New workforce entrant 1** | | |
| Blank | 0.0000 | 0.0000 |
| I made my own selection independent of anyone else | 0.0922 | 0.1075 |
| I selected a fund recommended by someone else | 0.1336 | 0.1469 |
| I used my employer’s default fund | 0.2767 | 0.3206 |
| I used my existing fund | 0.0000 | 0.0000 |
| Other | 0.4976 | 0.4250 |
| | | |
| **New workforce entrant 2** | | |
| Blank | 0.0000 | 0.0000 |
| I made my own selection independent of anyone else | 0.2042 | 0.2291 |
| I selected a fund recommended by someone else | 0.0976 | 0.1032 |
| I used my employer’s default fund | 0.4097 | 0.4566 |
| I used my existing fund | 0.0424 | 0.0089 |
| Other | 0.2462 | 0.2022 |
| | | |
| **Older individual 1** | | |
| Blank | 0.0000 | 0.0000 |
| I made my own selection independent of anyone else | 0.2144 | 0.2225 |
| I selected a fund recommended by someone else | 0.1697 | 0.1660 |
| I used my employer’s default fund | 0.5305 | 0.5467 |
| I used my existing fund | 0.0000 | 0.0000 |
| Other | 0.0854 | 0.0648 |
| | | |
| **Older individual 2** | | |
| Blank | 0.0000 | 0.0000 |
| I made my own selection independent of anyone else | 0.3113 | 0.3351 |
| I selected a fund recommended by someone else | 0.1367 | 0.1386 |
| I used my employer’s default fund | 0.4069 | 0.4351 |
| I used my existing fund | 0.0396 | 0.0080 |
| Other | 0.1055 | 0.0831 |

**a** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **b** 1998 observations were used for this regression.

## A.7.2 Did the choice experiment influence product preferences?

In section 4 the Commission presented results on the product preferences given by respondents. One concern could be that respondents who have just undertaken an assisted nomination task where they were presented with returns and fees, might be more inclined to have a preference for fees and returns. To explore this concern, a series of ordered probit regressions was conducted, each regression of the form:

$$P(Product\ feature\ ranked \leq\ j) = \Phi(\theta_j - X\beta)$$

Where $j = 1,2,3,4$ are the possible product preference ranks, and $\theta_j$ are the latent variable cut-points. The observables ($X$) used for this regression included, age brackets, a gender indicator, income brackets, level of highest education and the binary survey effort measure. Table A.13 presents the predicted probabilities for each of the hypothetical individuals under the control and treatment group. Only predictions have been shown as only effect sizes are relevant to this concern.

Table A.13 **Predicted probabilities of product feature rankings[a,b,c,d]**
By control and treatment groups

| *Ranking* | *4* | | *3* | | *2* | | *1* | |
|---|---|---|---|---|---|---|---|---|
| | Control | Treatment | Control | Treatment | Control | Treatment | Control | Treatment |
| **Fees** | | | | | | | | |
| New workforce entrant 1 | 0.1111 | 0.1119 | 0.1845 | 0.1854 | 0.3753 | 0.3754 | 0.3291 | 0.3273 |
| New workforce entrant 2 | 0.1121 | 0.1129 | 0.1857 | 0.1866 | 0.3754 | 0.3755 | 0.3268 | 0.3250 |
| Older individual 1 | 0.0843 | 0.0849 | 0.1518 | 0.1527 | 0.3641 | 0.3646 | 0.3998 | 0.3978 |
| Older individual 2 | 0.0961 | 0.0968 | 0.1670 | 0.1679 | 0.3712 | 0.3715 | 0.3657 | 0.3638 |
| **Net returns** | | | | | | | | |
| New workforce entrant 1 | 0.1784 | 0.1494 | 0.2977 | 0.2743 | 0.3112 | 0.3259 | 0.2127 | 0.2504 |
| New workforce entrant 2 | 0.1075 | 0.0888 | 0.2276 | 0.2009 | 0.3373 | 0.3345 | 0.3276 | 0.3758 |
| Older individual 1 | 0.0903 | 0.0743 | 0.2031 | 0.1771 | 0.3350 | 0.3263 | 0.3716 | 0.4223 |
| Older individual 2 | 0.0434 | 0.0354 | 0.1162 | 0.0977 | 0.2765 | 0.2516 | 0.5640 | 0.6152 |
| **Choice of investment options** | | | | | | | | |
| New workforce entrant 1 | 0.1904 | 0.2178 | 0.3739 | 0.3875 | 0.2403 | 0.2245 | 0.1955 | 0.1702 |
| New workforce entrant 2 | 0.1560 | 0.1796 | 0.3484 | 0.3670 | 0.2595 | 0.2465 | 0.2361 | 0.2070 |
| Older individual 1 | 0.2658 | 0.3002 | 0.4002 | 0.4024 | 0.1977 | 0.1799 | 0.1363 | 0.1175 |
| Older individual 2 | 0.1886 | 0.2159 | 0.3728 | 0.3867 | 0.2413 | 0.2256 | 0.1973 | 0.1718 |
| **Member services** | | | | | | | | |
| New workforce entrant 1 | 0.3219 | 0.3374 | 0.3262 | 0.3265 | 0.1936 | 0.1869 | 0.1583 | 0.1492 |
| New workforce entrant 2 | 0.5847 | 0.6016 | 0.2606 | 0.2526 | 0.0951 | 0.0900 | 0.0596 | 0.0558 |
| Older individual 1 | 0.5117 | 0.5292 | 0.2909 | 0.2843 | 0.1189 | 0.1130 | 0.0785 | 0.0736 |
| Older individual 2 | 0.8259 | 0.8357 | 0.1226 | 0.1161 | 0.0331 | 0.0310 | 0.0185 | 0.0172 |

**a** Each of the treatment effect coefficients are not statistically significant. **b** 1 represents the highest rank and 4 the lowest rank. **c** Respondents with level of highest level of education categorised as 'Other' based on their open responses have been removed. **d** 1992 observations were used for this regression.

From an economically significant point of view, the shortlist appeared to slightly encourage a higher ranking of returns. This can be seen by noting that the probability of hypothetical individuals ranking returns as most important increases by approximately five percentage points for each hypothetical individual. However the shortlist effects are insignificant. Overall this suggests that our results on product preferences might slightly overestimate the rankings on returns and slightly underestimate the other rankings (or might not), but considering the magnitudes, this does not materially affect the story.

### A.7.3 Did the choice experiment influence information rankings?

In section 4 the Commission presented results on the information source preferences given by respondents. One concern could be that respondents who have just undertaken an assisted nomination task where they were presented with easily comparable information and key performance indicators (KPI), might be more inclined to have a preference for information in the form of comparison websites and KPI. To explore this concern, a series of ordered probit regressions was conducted, each regression of the form:

$$P(Information\ source\ ranked \leq\ j) = \Phi(\theta_j - X\beta)$$

Where $j = 1,2,3,4,5,6$ are the possible product preference ranks, and $\theta_j$ are the latent variable cut-points.

The observables ($X$) used for this regression included, age brackets, a gender indicator, income brackets, level of highest education and the binary survey effort measure. Table A.14 presents the predicted probabilities for each of the hypothetical individuals under the control and treatment group. Only predictions have been shown as only effect sizes are relevant to this concern.

Overall from an economically significant point of view, there did not appear to be any concerns. Most differences between control and treatment are well below 5 per cent. There did not appear to be any clear patterns in how the probabilities change from control to treatment.

Table A.14 **Predicted probabilities of ranking each score for an information source[a,b,c]**

By control and treatment groups

| *Ranking* | *6* | | *5* | | *4* | | *3* | | *2* | | *1* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ctrl. | Treat | Ctrl. | Treat | Ctrl. | Treat | Ctrl. | Treat | Ctrl. | Treat | Ctrl. | Treat |
| **KPIs** | | | | | | | | | | | | |
| New workforce entrant 1 | 0.17 | 0.18 | 0.23 | 0.24 | 0.20 | 0.20 | 0.21 | 0.20 | 0.13 | 0.12 | 0.07 | 0.06 |
| New workforce entrant 2 | 0.06 | 0.07 | 0.12 | 0.12 | 0.15 | 0.15 | 0.25 | 0.25 | 0.25 | 0.24 | 0.18 | 0.17 |
| Older individual 1 | 0.07 | 0.07 | 0.13 | 0.13 | 0.15 | 0.16 | 0.25 | 0.25 | 0.24 | 0.23 | 0.16 | 0.15 |
| Older individual 2 | 0.02 | 0.02 | 0.05 | 0.05 | 0.07 | 0.08 | 0.17 | 0.18 | 0.29 | 0.30 | 0.39 | 0.37 |
| **PDSs** | | | | | | | | | | | | |
| New workforce entrant 1 | 0.27 | 0.23 | 0.25 | 0.24 | 0.22 | 0.23 | 0.14 | 0.15 | 0.08 | 0.10 | 0.04 | 0.05 |
| New workforce entrant 2 | 0.21 | 0.18 | 0.23 | 0.21 | 0.23 | 0.24 | 0.16 | 0.18 | 0.11 | 0.12 | 0.05 | 0.06 |
| Older individual 1 | 0.14 | 0.12 | 0.19 | 0.17 | 0.23 | 0.23 | 0.20 | 0.21 | 0.15 | 0.17 | 0.08 | 0.10 |
| Older individual 2 | 0.10 | 0.09 | 0.15 | 0.13 | 0.22 | 0.20 | 0.22 | 0.23 | 0.20 | 0.22 | 0.12 | 0.14 |
| **Comparison websites** | | | | | | | | | | | | |
| New workforce entrant 1 | 0.14 | 0.14 | 0.13 | 0.13 | 0.21 | 0.21 | 0.19 | 0.19 | 0.18 | 0.17 | 0.16 | 0.15 |
| New workforce entrant 2 | 0.10 | 0.11 | 0.10 | 0.11 | 0.18 | 0.19 | 0.19 | 0.19 | 0.21 | 0.20 | 0.21 | 0.20 |
| Older individual 1 | 0.10 | 0.10 | 0.10 | 0.10 | 0.18 | 0.18 | 0.19 | 0.19 | 0.21 | 0.21 | 0.22 | 0.21 |
| Older individual 2 | 0.06 | 0.06 | 0.06 | 0.07 | 0.13 | 0.14 | 0.17 | 0.18 | 0.24 | 0.24 | 0.33 | 0.32 |
| **Financial adviser** | | | | | | | | | | | | |
| New workforce entrant 1 | 0.11 | 0.10 | 0.08 | 0.08 | 0.16 | 0.15 | 0.16 | 0.16 | 0.18 | 0.18 | 0.32 | 0.32 |
| New workforce entrant 2 | 0.12 | 0.11 | 0.09 | 0.09 | 0.16 | 0.16 | 0.16 | 0.16 | 0.17 | 0.17 | 0.29 | 0.30 |
| Older individual 1 | 0.13 | 0.13 | 0.10 | 0.10 | 0.17 | 0.17 | 0.16 | 0.16 | 0.17 | 0.17 | 0.27 | 0.27 |
| Older individual 2 | 0.19 | 0.19 | 0.13 | 0.13 | 0.20 | 0.19 | 0.16 | 0.16 | 0.14 | 0.14 | 0.19 | 0.19 |
| **Friends or family** | | | | | | | | | | | | |
| New workforce entrant 1 | 0.11 | 0.11 | 0.17 | 0.17 | 0.14 | 0.14 | 0.16 | 0.16 | 0.19 | 0.18 | 0.24 | 0.23 |
| New workforce entrant 2 | 0.21 | 0.21 | 0.24 | 0.24 | 0.15 | 0.15 | 0.14 | 0.14 | 0.13 | 0.13 | 0.13 | 0.13 |
| Older individual 1 | 0.25 | 0.25 | 0.26 | 0.26 | 0.15 | 0.15 | 0.13 | 0.13 | 0.11 | 0.11 | 0.11 | 0.10 |
| Older individual 2 | 0.37 | 0.37 | 0.28 | 0.28 | 0.13 | 0.13 | 0.09 | 0.09 | 0.07 | 0.07 | 0.06 | 0.06 |
| **Employer** | | | | | | | | | | | | |
| New workforce entrant 1 | 0.13 | 0.14 | 0.19 | 0.20 | 0.16 | 0.16 | 0.13 | 0.13 | 0.18 | 0.18 | 0.21 | 0.20 |
| New workforce entrant 2 | 0.24 | 0.26 | 0.26 | 0.26 | 0.16 | 0.16 | 0.11 | 0.10 | 0.12 | 0.11 | 0.11 | 0.10 |
| Older individual 1 | 0.25 | 0.27 | 0.26 | 0.27 | 0.16 | 0.16 | 0.11 | 0.10 | 0.12 | 0.11 | 0.11 | 0.10 |
| Older individual 2 | 0.43 | 0.46 | 0.27 | 0.27 | 0.12 | 0.11 | 0.07 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 |

**a** 1 represents the highest rank and 6 the lowest rank. **b** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **c** 1992 observations were used for this regression.

# A.8 Regressions for nominations in the choice experiment

## A.8.1 Main regression

This section documents the regression associated with figure 23 in section 5 in the main text. The Commission is interested in exploring if the treatment can improve nomination rates and the role that demographics play. The regression equation is:

$$P(Nominates) = \Phi\left(\alpha + \beta_T T_i + \beta_{T,F} T_i f + \beta_f f + \beta_d D\right)$$

The results from this regression are presented in table A.16. A likelihood ratio test showed that the model is significant (table A.15).

Table A.15 **Log likelihood ratio test**
Main regression

| Quantity | *Value* | Degrees of freedom |
|---|---|---|
| *Null Deviance* | 308.18 | 1068 |
| *Residual Deviance* | 241.26 | 1044 |
| *Test Statistic* | 66.92 | |
| *Critical value – Chi-squared at 5% significance* | 37.65 | 25 |
| | | |
| **Conclusion** | **Model significant** | |

A supporting regression was conducted without interaction effects (table A.17).

## Table A.16 **Nomination rates in the choice experiment**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 0.6065 | 0.4768 | |
| **Treatment** | 0.1197 | 0.3861 | |
| **Survey effort** | 0.2269 | 0.3183 | |
| **Currently have fund** | 0.5063 | 0.2902 | * |
| **Age bracket** | | | |
| 25-34 | -0.5054 | 0.3215 | |
| 35-44 | -0.0670 | 0.3652 | |
| 45-54 | -0.2091 | 0.3348 | |
| 55-64 | -0.5875 | 0.3138 | * |
| **Female** | 0.1624 | 0.1932 | |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.6274 | 0.3125 | ** |
| 40 000 – 59 999 | 0.6303 | 0.3282 | * |
| 60 000 – 79 999 | 0.5682 | 0.3140 | * |
| 80 000 – 99 999 | 0.5763 | 0.3580 | |
| 100 000 – 129 999 | 1.0750 | 0.4526 | ** |
| 130 000 – 199 999 | 0.6822 | 0.4313 | |
| > 200 000 | 4.8201 | 325.2144 | |
| **Financial literacy score** | | | |
| 1 | -0.3609 | 0.3928 | |
| 2 | -0.3920 | 0.3994 | |
| 3 | 4.2912 | 279.2403 | |
| 4 | 0.2282 | 0.5788 | |
| **Treatment and financial literacy interactions** | | | |
| Treatment × 1 | 0.4115 | 0.5003 | |
| Treatment × 2 | 0.4550 | 0.5183 | |
| Treatment × 3 | -3.5635 | 279.2405 | |
| Treatment × 4 | -0.0401 | 0.6494 | |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.4976 | 0.3062 | |
| Certificate/Diploma | 0.2320 | 0.2859 | |
| Bachelor | -0.1135 | 0.3510 | |
| Graduate Diploma | 4.1407 | 318.2388 | |
| Postgraduate | 0.0246 | 0.4484 | |

**a** The constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, is in the control group, exhibited low survey effort and does not currently have a fund. **b** The education level ‘less than year 12’ is ‘primary school’, ‘some high school’, and ‘year 10 or equivalent’ combined. The education level ‘cert/dip’ is ‘trade certificate or apprenticeship’ and ‘diploma or associate degree’ combined. **c** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **d** 1069 matched observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

Table A.17 **Nomination rates in the choice experiment – no interactions**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 0.4758 | 0.4265 | |
| **Treatment** | 0.3419 | 0.1790 | * |
| **Survey effort** | 0.2069 | 0.3156 | |
| **Currently have fund** | 0.5157 | 0.2869 | * |
| **Age bracket** | | | |
| 25-34 | -0.5005 | 0.3166 | |
| 35-44 | -0.0404 | 0.3625 | |
| 45-54 | -0.2100 | 0.3304 | |
| 55-64 | -0.5714 | 0.3083 | * |
| **Female** | 0.1848 | 0.1912 | |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.6467 | 0.3093 | ** |
| 40 000 – 59 999 | 0.6113 | 0.3257 | * |
| 60 000 – 79 999 | 0.5567 | 0.3108 | * |
| 80 000 – 99 999 | 0.5263 | 0.3516 | |
| 100 000 – 129 999 | 1.0429 | 0.4456 | ** |
| 130 000 – 199 999 | 0.6927 | 0.4295 | |
| > 200 000 | 4.5530 | 203.8749 | |
| **Financial literacy score** | | | |
| 1 | -0.1183 | 0.2519 | |
| 2 | -0.1404 | 0.2820 | |
| 3 | 0.8055 | 0.4064 | ** |
| 4 | 0.1626 | 0.3482 | |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.4984 | 0.3030 | |
| Certificate/Diploma | 0.2496 | 0.2829 | |
| Bachelor | -0.1085 | 0.3456 | |
| Graduate Diploma | 3.9137 | 199.6699 | |
| Postgraduate | 0.0733 | 0.4481 | |

**a** The constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, is in the control group, exhibited low survey effort and does not currently have a fund. **b** The education level ‘less than year 12’ is ‘primary school’, ‘some high school’, and ‘year 10 or equivalent’ combined. The education level ‘cert/dip’ is ‘trade certificate or apprenticeship’ and ‘diploma or associate degree’ combined. **c** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **d** 1069 matched observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

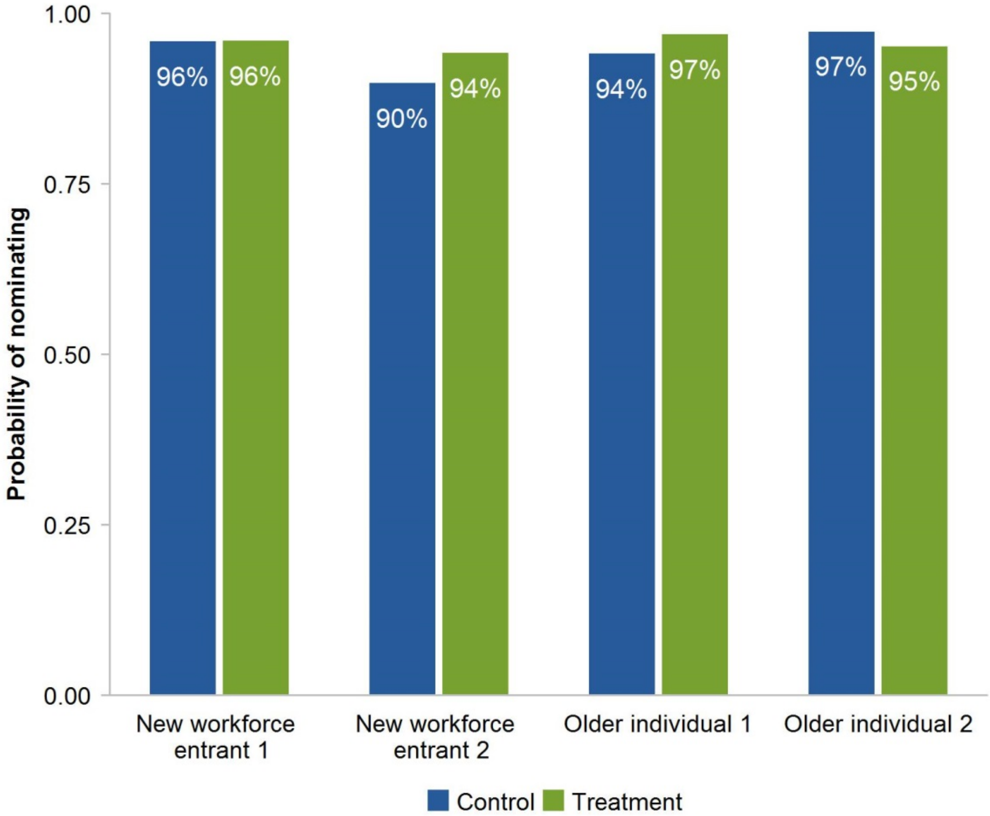
## A.8.2 Exploring the impact of nomination skippers

As highlighted in box 2 of section 5, a difficulty encountered in analysis were respondents in the assisted choice experiment who skipped the nomination task knowingly or not. The Commission decided to ignore these responses as there are many judgment calls to be made about whether a given respondent actually intended on nominating or not. In many cases it is unclear. This section explores some of the potential implications arising from this issue.

Regarding analysis about nomination rates, the Commission has conducted supporting analysis using data augmented with these skipping nomination respondents to get a sense of the possible range of values that the true treatment effect might take. It is important to emphasise that neither dataset (with or without the skipping nomination respondents) is likely to allow us to estimate the true treatment effect. Without the skipping nomination respondents, treatment effect estimates are likely to be too large. In figure 23 of section 5 it can be seen that treatment respondents are nominating at rates of 95 per cent and above. On the other hand in the group of nomination skippers, a conservative estimate places only 70 per cent of respondents as nominating. Thus pooling the groups together means the treatment effect is likely to be diminished. On the other hand, using the skipping nomination respondents might lead estimates being too small, because many respondents who may have intended on nominating have been classed as not nominating, due to our conservative approach.

Figure A.9 presents regression results analogous to figure 23 in section 5, using respondents who skipped their nominations and conservatively classified as nominating. The key change in figure A.9 (relative to figure 23) is that the treatment effects have been reduced in magnitude, and for older individual 2, reversed in sign. Taken together, the figures suggest that for some types of respondents, there are likely to be positive treatment effects on nomination rates, while for other demographics the effects will be minimal. Table A.18 presents the corresponding regression output.

Figure A.9 **Nomination rate predictions[a,b]**
Augmented dataset




**a** New workforce entrant 2 in this case has been specified with a financial literacy score of 2 instead of 3. This is due to there being coefficients associated with a financial literacy score of 3 having very large standard errors. **b** Unassisted choice selections were manually processed by inspection and bucketing. Cases where respondents have 'attempted' to nominate have been classified as respondents nominating.

The effects on other choice experiment analyses should be less of an issue. For example, the average age and financial literacy score of those skipping nominations are 37 and 1.85 versus those who nominate, which are 41 and 2.1 respectively. Thus there are only small differences between those who skipped and those who did not. These differences may be brought about by the 30 per cent who skipped and clearly intended on not nominating. Therefore the implications of these issues on our analysis where nomination rates are not concerned are relatively minor. Matching methods will mitigate the risks associated with removing these respondents on achieving balance and maintaining representativeness in analysis. The power simulations outlined in section A.5.3 were conducted taking into account this issue.

Table A.18 **Nomination rates in the choice experiment – with nomination skippers**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 0.6079 | 0.3859 | |
| **Treatment** | -0.4549 | 0.3025 | |
| **Survey effort** | 0.6969 | 0.2126 | *** |
| **Currently have fund** | 0.6151 | 0.2291 | *** |
| **Age bracket** | | | |
| 25-34 | -0.6780 | 0.2489 | *** |
| 35-44 | -0.3516 | 0.2611 | |
| 45-54 | -0.3144 | 0.2631 | |
| 55-64 | -0.5814 | 0.2595 | ** |
| **Female** | 0.0296 | 0.1429 | |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.4869 | 0.2688 | * |
| 40 000 – 59 999 | 0.2510 | 0.2603 | |
| 60 000 – 79 999 | 0.2626 | 0.2588 | |
| 80 000 – 99 999 | 0.1966 | 0.2930 | |
| 100 000 – 129 999 | 0.4387 | 0.3095 | |
| 130 000 – 199 999 | 0.0781 | 0.3109 | |
| > 200 000 | 0.5914 | 0.4656 | |
| **Financial literacy score** | | | |
| 1 | -0.1929 | 0.3520 | |
| 2 | -0.2454 | 0.3561 | |
| 3 | 4.0730 | 109.6598 | |
| 4 | 0.4465 | 0.5338 | |
| **Treatment and financial literacy interactions** | | | |
| Treatment × 1 | 0.4623 | 0.4042 | |
| Treatment × 2 | 0.7577 | 0.4208 | * |
| Treatment × 3 | -2.9878 | 109.6601 | |
| Treatment × 4 | 0.1831 | 0.5705 | |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.1396 | 0.2373 | |
| Certificate/Diploma | -0.0213 | 0.2327 | |
| Bachelor | -0.0712 | 0.2723 | |
| Graduate Diploma | 0.6455 | 0.5491 | |
| Postgraduate | -0.2234 | 0.3134 | |

**a** The constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, is in the control group, exhibited low survey effort and does not currently have a fund. **b** The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined. **c** Respondents with level of highest level of education categorised as 'Other' based on their open responses have been removed. **d** 1192 matched observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

## A.9 Regression for nominations across shortlists

This section documents the regression associated with section 6 in the main text. The Commission is interested in exploring if varying the number of options presented and/or varying the presentation of fees and returns in a percentage or dollar form can influence nomination rates. The regression equation is:

$$P(Nominates) = \Phi\left(\alpha + \beta_T T + \beta_P P + \beta_f f + \beta_d D\right)$$

Where $T$ refers to the number of options presented with associated coefficient vector $\beta_T$ and $P$ is an indicator for whether the information was presented in a percentage or dollar form, with corresponding coefficient $\beta_P$.The results from this regression are presented in table A.20. A likelihood ratio test showed that the model is significant (table A.19).

Table A.19 **Log likelihood ratio test**

Nomination rates and shortlist design

| Quantity | *Value* | Degrees of freedom |
|---|---|---|
| *Null Deviance* | 328.64 | 1425 |
| *Residual Deviance* | 277.49 | 1397 |
| *Test Statistic* | 51.15 | |
| *Critical value – Chi-squared at 5 per cent significance* | 42.56 | 29 |
| | | |
| **Conclusion** | **Model significant** | |

## Table A.20 **Nomination rates and shortlist design**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | 0.9699 | 0.4248 | ** |
| **Number of options** | | | |
| 5 | 0.1561 | 0.2431 | |
| 6 | 0.2522 | 0.2442 | |
| 7 | 0.1374 | 0.2373 | |
| 8 | 0.3947 | 0.2546 | |
| **Per cent** | -0.1651 | 0.1611 | |
| **Survey effort** | 0.2583 | 0.3026 | |
| **Currently have fund** | 0.1274 | 0.1979 | |
| **Age bracket** | | | |
| 25-34 | -0.1276 | 0.2966 | |
| 35-44 | -0.4097 | 0.2872 | |
| 45-54 | -0.2563 | 0.2923 | |
| 55-64 | -0.5791 | 0.2720 | ** |
| **Female** | 0.0954 | 0.1701 | |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.0922 | 0.2684 | |
| 40 000 – 59 999 | 0.2426 | 0.2842 | |
| 60 000 – 79 999 | 0.1393 | 0.2830 | |
| 80 000 – 99 999 | 0.4426 | 0.3319 | |
| 100 000 – 129 999 | 0.7510 | 0.4025 | * |
| 130 000 – 199 999 | 0.4452 | 0.3867 | |
| > 200 000 | 4.2161 | 172.5934 | |
| **Financial literacy score** | | | |
| 1 | 0.2306 | 0.2202 | |
| 2 | 0.3192 | 0.2285 | |
| 3 | 1.1159 | 0.3665 | *** |
| 4 | 0.5986 | 0.3090 | ** |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.5617 | 0.2878 | * |
| Certificate/Diploma | 0.2224 | 0.2263 | |
| Bachelor | 0.1038 | 0.2465 | |
| Graduate Diploma | -0.0972 | 0.3322 | |
| Postgraduate | 0.5557 | 0.4745 | |

[a] The constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, was presented with 4 options and information in dollars, exhibited low survey effort and does not currently have a fund. [b] The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined. [c] Respondents with level of highest level of education categorised as 'Other' based on their open responses have been removed. [d] 1426 observations were used for this regression. [e] ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

## A.10 Regression for difficulty across control and treatment

In figure 27 in section 6, the Commission presented results on regression analysis about difficulty, and how having a shortlist might affect the difficulty. The regression equation is

$$P(Difficulty \leq j) = \Phi\left(\theta_j - \left(\alpha + \beta_T T + \beta_{T,F} Tf + \beta_f f + \beta_d D\right)\right)$$

Where $j = 1{,}2{,}3{,}4{,}5$ are the possible difficulty scores, with 5 being the most difficulty, and $\theta_j$ are the latent variable cut-points. The results from this regression are presented in table A.22. A likelihood ratio test showed that the model is significant (table A.21).

Table A.21 **Log likelihood ratio test**
Difficulty across treatment and control

| Quantity | *Value* | Degrees of freedom |
| --- | --- | --- |
| *Null Deviance* | 3034.69 | 1065 |
| *Residual Deviance* | 2929.57 | 1033 |
| *Test Statistic* | 105.12 | |
| *Critical value – Chi-squared at 5% significance* | 46.19 | 32 |
| | | |
| **Conclusion** | **Model significant** | |

To explore the treatment effects a bit further, the Commission conducted a similar regression, but with no interaction effects (table A.23). The results showed that the treatment and financial literacy scores gain significance, again suggesting a potential lack of data for identifying interaction effects.

## Table A.22 **Difficulty across treatment and control**[a,b,c,d,e]

| *Variable* | | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|---|
| **Cut-points:** | 1\|2 | -0.5094 | 0.2540 | ** |
| | 2\|3 | 0.0769 | 0.2538 | |
| | 3\|4 | 0.7550 | 0.2542 | *** |
| | 4\|5 | 1.4945 | 0.2580 | *** |
| **Treatment** | | 0.1513 | 0.1992 | |
| **Survey effort** | | -0.4370 | 0.1533 | *** |
| **Currently have fund** | | -0.4003 | 0.1291 | *** |
| **Age bracket** : | 25-34 | -0.0481 | 0.1349 | |
| | 35-44 | -0.0729 | 0.1307 | |
| | 45-54 | -0.2279 | 0.1326 | * |
| | 55-64 | -0.2206 | 0.1353 | |
| **Female** | | 0.1650 | 0.0728 | ** |
| **Household income bracket ($)** | | | | |
| 20 000 – 39 999 | | 0.2314 | 0.1531 | |
| 40 000 – 59 999 | | 0.1420 | 0.1549 | |
| 60 000 – 79 999 | | 0.0293 | 0.1544 | |
| 80 000 – 99 999 | | -0.0262 | 0.1642 | |
| 100 000 – 129 999 | | -0.1022 | 0.1656 | |
| 130 000 – 199 999 | | -0.1711 | 0.1747 | |
| > 200 000 | | 0.0737 | 0.2180 | |
| **Financial literacy score:** | 1 | 0.2788 | 0.2147 | |
| | 2 | 0.1684 | 0.2183 | |
| | 3 | -0.1256 | 0.2208 | |
| | 4 | -0.2728 | 0.2460 | |
| **Treatment and financial literacy interactions** | | | | |
| Treatment × 1 | | -0.1118 | 0.2566 | |
| Treatment × 2 | | 0.0900 | 0.2608 | |
| Treatment × 3 | | 0.1678 | 0.2551 | |
| Treatment × 4 | | 0.2277 | 0.2771 | |
| **Level of highest education** | | | | |
| Year 12 or equivalent | | 0.3748 | 0.1426 | *** |
| Certificate/Diploma | | 0.1623 | 0.1412 | |
| Bachelor | | 0.5325 | 0.1593 | *** |
| Graduate Diploma | | 0.4668 | 0.2150 | ** |
| Postgraduate | | 0.5835 | 0.1803 | *** |

**a** Each constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, is in the control group, exhibited low survey effort and does not currently have a fund. **b** The education level ‘less than year 12’ is ‘primary school’, ‘some high school’, and ‘year 10 or equivalent’ combined. The education level ‘cert/dip’ is ‘trade certificate or apprenticeship’ and ‘diploma or associate degree’ combined. **c** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **d** 1069 matched observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

Table A.23 **Difficulty across treatment and control – no interactions**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Cut-points** | | | |
| 1\|2 | -0.4511 | 0.2223 | ** |
| 2\|3 | 0.1343 | 0.2221 | |
| 3\|4 | 0.8119 | 0.2225 | *** |
| 4\|5 | 1.5507 | 0.2269 | *** |
| **Treatment** | 0.2223 | 0.0775 | *** |
| **Survey effort** | -0.4308 | 0.1533 | *** |
| **Currently have fund** | -0.4021 | 0.1286 | *** |
| **Age bracket** | | | |
| 25-34 | -0.0387 | 0.1345 | |
| 35-44 | -0.0690 | 0.1303 | |
| 45-54 | -0.2139 | 0.1317 | |
| 55-64 | -0.2074 | 0.1347 | |
| **Female** | 0.1611 | 0.0725 | ** |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.2244 | 0.1527 | |
| 40 000 – 59 999 | 0.1494 | 0.1547 | |
| 60 000 – 79 999 | 0.0244 | 0.1541 | |
| 80 000 – 99 999 | -0.0152 | 0.1638 | |
| 100 000 – 129 999 | -0.0991 | 0.1654 | |
| 130 000 – 199 999 | -0.1722 | 0.1745 | |
| > 200 000 | 0.0888 | 0.2173 | |
| **Financial literacy score** | | | |
| 1 | 0.1955 | 0.1203 | |
| 2 | 0.2247 | 0.1279 | * |
| 3 | -0.0076 | 0.1259 | |
| 4 | -0.1054 | 0.1375 | |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.3803 | 0.1424 | *** |
| Certificate/Diploma | 0.1620 | 0.1410 | |
| Bachelor | 0.5345 | 0.1590 | *** |
| Graduate Diploma | 0.4614 | 0.2148 | ** |
| Postgraduate | 0.5765 | 0.1802 | *** |

**a** Each constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, is in the control group, exhibited low survey effort and does not currently have a fund. **b** The education level 'less than year 12' is 'primary school', 'some high school', and 'year 10 or equivalent' combined. The education level 'cert/dip' is 'trade certificate or apprenticeship' and 'diploma or associate degree' combined. **c** Respondents with level of highest level of education categorised as 'Other' based on their open responses have been removed. **d** 1069 matched observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

## A.11 Regression for difficulty by shortlist design

In section 6, the Commission considered the influence shortlist design might have on difficulty. Supporting regression analysis was conducted, and this section presents those results.

To assess the influence the number of options might have on difficulty the commission conducted a regression (without matching[21]) with equation:

$$P(Difficulty \leq\ j) = \Phi\left(\theta_j - \left(\alpha + \beta_T T + \beta_f f + \beta_d D\right)\right)$$

Where $j = 1,2,3,4,5$ are the possible difficulty scores, with five being the most difficuly, $\theta_j$ are the latent variable cut-points and in this case $T$ represents the number of options presented to respondents with $\beta_T$ being the corresponding vector of coefficients. The results from this regression are presented in table A.25. A likelihood ratio test showed that the model is significant (table A.24).

Table A.24 **Log likelihood ratio test**

Difficulty across number of options presented

| Quantity | *Value* | Degrees of freedom |
|---|---|---|
| *Null Deviance* | 4116.61 | 1422 |
| *Residual Deviance* | 3969.79 | 1395 |
| *Test Statistic* | 146.82 | |
| *Critical value – Chi-squared at 5% significance* | 44.99 | 31 |
| | | |
| **Conclusion** | **Model significant** | |

To assess how the presentation of information might influence difficulty the commission conducted a regression (with matching) with equation:

$$P(Difficulty \leq\ j) = \Phi\left(\theta_j - \left(\alpha + \beta_P P + \beta_f f + \beta_d D\right)\right)$$

Where $j = 1,2,3,4,5$ are the possible difficulty scores, with 5 being the most difficuly, $\theta_j$ are the latent variable cut-points and $P$ is an indicator for whether information was presented in per cent, with associated coefficient $\beta_P$. The results from this regression are presented in table A.26. A likelihood ratio test showed that the model is significant (table A.27).

21 Matching across multiple treatment groups would severely reduce sample sizes.

## Table A.25 Difficulty by number of options presented[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Cut-points** | | | |
| 1\|2 | -1.0504 | 0.1910 | *** |
| 2\|3 | -0.4298 | 0.1899 | ** |
| 3\|4 | 0.3145 | 0.1901 | * |
| 4\|5 | 0.9721 | 0.1930 | *** |
| **Number of options presented** | | | |
| 5 | 0.0094 | 0.0924 | |
| 6 | 0.0286 | 0.0918 | |
| 7 | 0.0762 | 0.0940 | |
| 8 | 0.1462 | 0.0905 | |
| **Survey effort** | -0.7153 | 0.1360 | *** |
| **Currently have fund** | -0.1207 | 0.0830 | |
| **Age bracket** | | | |
| 25-34 | -0.2307 | 0.0994 | ** |
| 35-44 | -0.3463 | 0.1005 | *** |
| 45-54 | -0.5821 | 0.1016 | *** |
| 55-64 | -0.5199 | 0.1014 | *** |
| **Female** | 0.2127 | 0.0607 | *** |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.1170 | 0.1264 | |
| 40 000 – 59 999 | 0.0971 | 0.1270 | |
| 60 000 – 79 999 | 0.1614 | 0.1301 | |
| 80 000 – 99 999 | 0.1103 | 0.1316 | |
| 100 000 – 129 999 | -0.0776 | 0.1395 | |
| 130 000 – 199 999 | 0.2462 | 0.1733 | |
| > 200 000 | 0.1170 | 0.1264 | |
| **Financial literacy score** | | | |
| 1 | -0.0634 | 0.1016 | |
| 2 | 0.0080 | 0.0994 | |
| 3 | -0.1100 | 0.1022 | |
| 4 | -0.2041 | 0.1140 | * |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.1403 | 0.1045 | |
| Certificate/Diploma | 0.1616 | 0.0999 | |
| Bachelor | 0.2830 | 0.1043 | *** |
| Graduate Diploma | 0.3452 | 0.1440 | ** |
| Postgraduate | 0.3562 | 0.1299 | *** |

**a** Each constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, exhibited low survey effort, does not currently have a fund and who was presented with 4 options. **b** The education level ‘less than year 12’ is ‘primary school’, ‘some high school’, and ‘year 10 or equivalent’ combined. The education level ‘cert/dip’ is ‘trade certificate or apprenticeship’ and ‘diploma or associate degree’ combined. **c** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **d** 1426 observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

## Table A.26 **Difficulty by presentation of information**[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Cut-points** | | | |
| 1\|2 | -1.0010 | 0.2363 | *** |
| 2\|3 | -0.3466 | 0.2350 | |
| 3\|4 | 0.4115 | 0.2352 | * |
| 4\|5 | 1.0561 | 0.2393 | *** |
| **Per cent** | -0.0490 | 0.0725 | |
| **Survey effort** | -0.5323 | 0.1782 | *** |
| **Currently have fund** | -0.0437 | 0.1260 | |
| **Age bracket** | | | |
| 25-34 | -0.2006 | 0.1348 | |
| 35-44 | -0.3567 | 0.1363 | *** |
| 45-54 | -0.5766 | 0.1384 | *** |
| 55-64 | -0.4565 | 0.1368 | *** |
| **Female** | 0.2102 | 0.0766 | *** |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | 0.2733 | 0.1645 | * |
| 40 000 – 59 999 | 0.1980 | 0.1662 | |
| 60 000 – 79 999 | 0.2184 | 0.1643 | |
| 80 000 – 99 999 | 0.2271 | 0.1715 | |
| 100 000 – 129 999 | 0.0817 | 0.1734 | |
| 130 000 – 199 999 | -0.0391 | 0.1852 | |
| > 200 000 | 0.5161 | 0.2313 | ** |
| **Financial literacy score** | | | |
| 1 | -0.0532 | 0.1256 | |
| 2 | -0.1247 | 0.1323 | |
| 3 | -0.1733 | 0.1318 | |
| 4 | -0.3425 | 0.1445 | *** |
| **Level of highest education** | | | |
| Year 12 or equivalent | 0.0607 | 0.1509 | |
| Certificate/Diploma | -0.0023 | 0.1427 | |
| Bachelor | 0.1641 | 0.1520 | |
| Graduate Diploma | 0.0725 | 0.2329 | |
| Postgraduate | 0.3453 | 0.1855 | * |

**a** Each constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, exhibited low survey effort, does not currently have a fund and who was presented with information in dollars. **b** The education level ‘less than year 12’ is ‘primary school’, ‘some high school’, and ‘year 10 or equivalent’ combined. The education level ‘cert/dip’ is ‘trade certificate or apprenticeship’ and ‘diploma or associate degree’ combined. **c** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **d** 939 matched observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

Table A.27 **Log likelihood ratio test**
Difficulty across presentation of information

| Quantity | *Value* | Degrees of freedom |
|---|---|---|
| *Null Deviance* | 2674.85 | 935 |
| *Residual Deviance* | 2589.21 | 911 |
| *Test Statistic* | 85.64 | |
| *Critical value – Chi-squared at 5% significance* | 41.34 | 28 |
| | | |
| **Conclusion** | **Model significant** | |

## A.12 Regression for heterogeneity in selections

This section documents the regression associated with section 7 in the main text. The Commission is interested in exploring factors leading to respondents choosing fund characteristic set one over fund characteristic set five. The regression equation is:

$$P(Chooses\ fund\ characteristic\ set\ 1\ over\ fund\ characteristic\ set\ 5) \\ = \Phi(\alpha + \beta_B B + \beta_D D)$$

Where $B$ refers to the fund-level fixed effects presented to respondents with associated coefficient vector $\beta_B$. The results from this regression are presented in table A.29. A likelihood ratio test showed that the model is significant (table A.28).

Table A.28 **Log likelihood ratio test**
Difficulty across number of options presented

| Quantity | *Value* | Degrees of freedom |
|---|---|---|
| *Null Deviance* | 304.38 | 729 |
| *Residual Deviance* | 257.76 | 699 |
| *Test Statistic* | 46.62 | |
| *Critical value – Chi-squared at 5% significance* | 44.99 | 31 |
| | | |
| **Conclusion** | **Model significant** | |

## Table A.29 Rates of choosing fund characteristic set 1 over fund characteristic set 5[a,b,c,d,e]

| *Variable* | *Coefficient* | *Standard Error* | *Significance* |
|---|---|---|---|
| **Constant** | -0.8193 | 0.5019 | |
| **Fund level fixed effects** | | | |
| Fund 2 | -0.0919 | 0.2634 | |
| Fund 3 | -0.4964 | 0.4032 | |
| Fund 4 | -0.5503 | 0.3947 | |
| Fund 5 | -0.4639 | 0.3622 | |
| Fund 6 | -1.0492 | 0.5498 | * |
| Fund 7 | -0.0361 | 0.2867 | |
| Fund 8 | -0.3911 | 0.3698 | |
| **Survey effort** | -0.3504 | 0.3117 | |
| **Currently have fund** | 0.3760 | 0.2578 | |
| **Age bracket** | | | |
| 25-34 | 0.6134 | 0.3057 | ** |
| 35-44 | 0.3364 | 0.3283 | |
| 45-54 | 0.2343 | 0.3350 | |
| 55-64 | 0.3740 | 0.3334 | |
| **Female** | 0.0659 | 0.1856 | |
| **Household income bracket ($)** | | | |
| 20 000 – 39 999 | -0.1535 | 0.3568 | |
| 40 000 – 59 999 | -0.2811 | 0.3732 | |
| 60 000 – 79 999 | -0.4000 | 0.3732 | |
| 80 000 – 99 999 | 0.1172 | 0.3725 | |
| 100 000 – 129 999 | 0.0069 | 0.3708 | |
| 130 000 – 199 999 | -0.3767 | 0.4435 | |
| > 200 000 | 0.2387 | 0.4693 | |
| **Financial literacy score** | | | |
| 1 | -0.0541 | 0.2656 | |
| 2 | -0.1596 | 0.2593 | |
| 3 | -0.5365 | 0.2981 | * |
| 4 | -0.5909 | 0.3466 | * |
| **Level of highest education** | | | |
| Year 12 or equivalent | -0.6996 | 0.2835 | * |
| Certificate/Diploma | -0.8220 | 0.2733 | *** |
| Bachelor | -0.7999 | 0.2870 | *** |
| Graduate Diploma | -0.8404 | 0.4363 | * |
| Postgraduate | -0.3389 | 0.3316 | |

**a** The constant reflects a respondent who scored 0 on the financial and super literacy questions, in the 15-24 age bracket, is male, is in the < 20 000 income bracket, with a level of highest education which is less than year 12, exhibited low survey effort, does not currently have a fund and selected fund 1. **b** The education level ‘less than year 12’ is ‘primary school’, ‘some high school’, and ‘year 10 or equivalent’ combined. The education level ‘cert/dip’ is ‘trade certificate or apprenticeship’ and ‘diploma or associate degree’ combined. **c** Respondents with level of highest level of education categorised as ‘Other’ based on their open responses have been removed. **d** 730 observations were used for this regression. **e** ***, ** and * represent statistical significance at the 1, 5 and 10 per cent level respectively

# References

ABS (Australian Bureau of Statistics) 2009, Employment Arrangements, Retirement and Superannuation, Apr to Jul 2007, Cat. no. 6361.0, Canberra.

—— 2016, Education and Work, May 2016, Cat. no. 6227.0, Canberra.

Angrist, J.D. and Pischke, J.-S. 2009, Mostly Harmless Econometrics: An Empiricist's Companion.

Blair, G. and Imai, K. 2012, 'Statistical analysis of list experiments', Political Analysis, vol. 20, no. 1, pp. 47–77.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J. and Munafò, M.R. 2013, 'Power failure: why small sample size undermines the reliability of neuroscience', Nature Reviews Neuroscience, vol. 14, no. 5, pp. 365–376.

Colmar Brunton 2010, Investigating Superannuation: Quantitative Investigation with Superannuation Consumers, Final Quantitative Report, prepared for the Australian Taxation Office, Canberra.

GCA (Graduate Careers Australia) 2015, Graduate Salaries 2015: A report on the earnings of new Australian graduates in their first full-time employment.

Gelman, A. and Carlin, J. 2014, 'Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors', Perspectives on Psychological Science, vol. 9, no. 6, pp. 641–651.

Gordon, B., Zettelmeyer, F., Bhargava, N. and Chapsky, D. 2017, A Comparison of Approaches to Advertising Measurement; Evidence from Big Field Experiments at Facebook, Working Paper, 24 April.

Ho, D., Imai, K., King, G. and Stuart, E. 2007, 'Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference', Political Analysis, vol. 15, pp. 199–236.

Holbrook, A.L. and Krosnick, J.A. 2010, 'Social desirability bias in voter turnout reports: Tests using the item count technique', Public Opinion Quarterly, vol. 74, no. 1, pp. 37–67.

Iacus, S., King, G. and Porro, G. 2011, 'Causal inference without balance checking: Coarsened exact matching', Political Analysis.

King, G., Nielsen, R., Coberly, C., Pope, J.E. and Aaron, W. 2011, Comparative Effectiveness of Matching Methods for Causal Inference, Working Paper.

Lusardi, A. and Mitchell, O.S. 2011, 'Financial literacy and retirement planning in the United States', Journal of Pension Economics and Finance, vol. 10, pp. 509–525.

Minifie, J., Cameron, T. and Savage, J. 2014, Super Sting: How to Stop Australians Paying Too Much for Superannuation, Grattan Institute, Melbourne.

OECD (Organisation for Economic Co-operation and Development) 2011, Measuring Financial Literacy: Questionnaire and Guidance Notes for conducting an Internationally Comparable Survey of Financial Literacy, International Network on Financial Education, Paris.

Ortmann, A. 2015, 'The replication crisis has engulfed economics', The Conversation, http://theconversation.com/the-replication-crisis-has-engulfed-economics-49202 (accessed 3 May 2017).

PC (Productivity Commission) 2017, Superannuation: Alternative Default Models, Inquiry Draft Report, March, Canberra.

Rubin, D.B. 2008, 'For objective causal inference, design trumps analysis', Annals of Applied Statistics, vol. 2, no. 3, pp. 808–840.

Russ, R. and Gelman, A. 2017, 'Andrew Gelman on social science, small samples, and the garden of the forking paths', EconTalk, http://www.econtalk.org/archives/2017/03/andrew_gelman_o.html (accessed 3 May 2017).

Sniderman, P.M., Tetlock, P.E. and Piazza, T. 1992, Codebook for the 1991 National Race and Politics Survey.