

Gold Natalie (Orcid ID: 0000-0003-0706-1618)

Burton Robyn (Orcid ID: 0000-0003-1684-5238)

Arambepola Rohan (Orcid ID: 0000-0003-2833-8786)

**Effect of alcohol label designs with different pictorial representations of alcohol content and health warnings on knowledge and understanding of Low Risk Drinking Guidelines: A randomized controlled trial**

Natalie Gold,<sup>1,2</sup> Mark Egan,<sup>3</sup> Kristina Londakova,<sup>3</sup> Abigail Mottershaw,<sup>3</sup> Hugo Harper,<sup>3</sup> Robyn Burton,<sup>1,4</sup> Clive Henn,<sup>1</sup> Maria Smolar,<sup>1</sup> Matthew Walmsley,<sup>1</sup> Rohan Arambepola,<sup>1,5,6</sup> Robin Watson,<sup>1,7</sup> Sarah Bowen,<sup>1,8</sup> Felix Greaves,<sup>1,9</sup>

<sup>1</sup> Public Health England, Wellington House, 133-155 Waterloo Road, London SE1 8UG

<sup>2</sup> Department of Philosophy, University of Oxford, UK, Radcliffe Humanities, Woodstock Road, Oxford OX2 6GG

<sup>3</sup> Behavioural Insights Team, 4 Matthew Parker St, Westminster, London SW1H 9NP

<sup>4</sup> Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London SE5 8AF

<sup>5</sup> Oxford Big Data Institute, Old Road Campus, Oxford OX3 7LF

<sup>6</sup> Nuffield Department of Medicine, University of Oxford, Richard Doll Building, Old Road Campus, Oxford OX3 7LF

<sup>7</sup> Department of Anthropology, Durham University, Dawson Building, South Road, Durham, DH1 3LE

<sup>8</sup> School of Economics, Sir Clive Granger Building University Park Nottingham NG7 2RD UK

<sup>9</sup> Department of Primary Care and Public Health, Imperial College, London, UK South Kensington, London SW7 2AZ

Correspondence to:

Dr Natalie Gold, Public Health England

**[natalie.gold@phe.gov.uk](mailto:natalie.gold@phe.gov.uk)**

**Conflict of interest declaration**

There are no conflicts of interest to be declared

**Running head:** Alcohol label designs and knowledge of LRDGs

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/add.15327

## Abstract

### *Background and aims:*

The UK Low Risk Drinking Guidelines (LRDG) recommend not regularly drinking more than 14 units of alcohol per week. We tested the effect of different pictorial representations of alcohol content, some with a health warning, on knowledge of the LRDG and understanding of how many drinks it equates to.

### *Design:*

Parallel randomized controlled trial.

### *Setting:*

Online, 25 Jan - 1 Feb 2019.

### *Participants:*

Participants ( $n = 7,516$ ) were English, over 18 years, and drink alcohol.

### *Interventions:*

The control group saw existing industry-standard labels; six intervention groups saw designs based on: food labels (serving or serving & container), pictographs (servings or containers), pie charts (servings), or risk gradients. A total of 500 participants (~70 per condition) saw a health warning under the design.

### *Measurements:*

Primary outcomes: (i) knowledge: proportion who answered that the LRDG is 14 units; (ii) understanding: how many servings/ containers of beverages one can drink before reaching 14 units (10 questions, average distance from correct answer).

### *Findings:*

In the control group, 21.5% knew the LRDG; proportions were higher in intervention groups (all  $p < 0.001$ ). The three best-performing designs had the LRDG in a separate statement, underneath the pictograph container, 51.1% (AOR = 3.74, 95% CI 3.08-4.54), pictograph serving 48.8% (AOR = 4.11, 95% CI 3.39-4.99), and pie chart serving, 47.5% (AOR = 3.57, 95% CI 2.93-4.34). Participants underestimated how many servings they could drink: control  $M = -4.64$ ,  $SD = 3.43$ ; intervention groups were more accurate (all  $p < 0.001$ ), best performing was pictograph serving ( $M = -0.93$ ,  $SD = 3.43$ ). Participants overestimated how many containers they could drink: control  $M = 0.09$ ,  $SD = 1.02$ ; intervention groups overestimated even more (all  $p < 0.007$ ), worst performing was food label serving ( $M = 1.10$ ,  $SD = 1.27$ ). Participants judged the alcohol content of beers more accurately than wine or spirits. The inclusion of a health warning had no statistically significant effect on any measure.

### *Conclusions:*

Labels with enhanced pictorial representations of alcohol content improved knowledge and understanding of the United Kingdom's Low Risk Drinking Guidelines (LRDG) compared with industry-standard labels; health warnings did not improve knowledge or understanding of LRDG. Designs that improved knowledge most had the LRDG in a separate statement located underneath graphics.

**Keywords**

Alcohol, Alcohol unit, Cancer, Consumer knowledge, Graphic labels, Health warning label, Labelling, Low risk drinking guidelines, Pictorial labels, Product labelling, Standard drink

## Introduction

Alcohol consumption is associated with over 200 diseases, injuries, and conditions (1). For all conditions there is a dose-response relationship – with increasing levels of alcohol consumption there is increasing risk (2). For some conditions, such as liver disease, this relationship is exponential (3), whereas for other conditions, such as some cancers, it is linear (4). The most effective way of reducing these risks is reducing individual- and population-level consumption (5). As such, many governments have developed low risk drinking guidelines (LRDGs), which commonly include a recommended daily or weekly maximum intake, expressed as numbers of “standard drinks” or “units of alcohol” (6, 7). The World Health Organisation defines a standard drink as 10g of pure ethanol and advises people not to exceed two standard drinks per day (8). Although widespread, LRDGs are not universal or uniform. A review of 37 government agency guidelines found that guidelines for low-risk consumption ranged from 10-56g of ethanol per day and that the standard drink sizes (which the guidelines were expressed in) ranged from 8-20g of ethanol, with 10g as the modal size (7).

The UK government published LRDGs in the 1990s (9, 10), which were updated by the United Kingdom Chief Medical Officers (CMOs) in 2016 (11). The weekly drinking guidelines for both men and women state, “to keep health risks from alcohol to a low level it is safest not to regularly drink more than 14 units a week on a regular basis” (11). More than 10 million adults in the UK drink more than the LRDG of 14 units per week (12). The Department of Health recommended that the CMO’s guidelines be communicated to the general public using visual prompts (13). In 2011, the Government in England launched the Public Health Responsibility Deal involving voluntary agreements with industry, which included labelling at least 80% of alcohol products with unit content, low-risk guidelines,

pregnancy warnings and responsibility statements (14). However, a market survey conducted in 2014 found that only 57% of labels met best practice as defined by the Portman Group (15).

The LRDG were developed on the principles that: (a) people have a right to accurate information and clear advice about alcohol and its health risks, and (b) government has a responsibility to ensure this information is provided for the public in a clear and open way, so it can make informed choices. However, there is a lack of knowledge and understanding of the LRDG. Recent representative surveys of the adult British population have found that only between 8% and 25% know that the LRDG is 14 units per week (16-18). Even where people know the guidelines, they may not understand them. The CMOs' guidelines use "units" of alcohol as a measure. A unit is 10ml or 8g of pure alcohol (about two teaspoons) (19).

However, research shows that people find it difficult to use units to gauge their alcohol intake, which is not surprising given that alcoholic drinks vary widely in their strengths and serving sizes (20). Further, knowledge of the harms that alcohol causes is poor. In a 2018 UK survey, in answer to an open response question about which health conditions can result from drinking alcohol, only 40% of respondents identified liver damage/failure as a drinking outcome and 31% reported cancer (17).

A review of the effectiveness of labelling approaches, where labels on alcohol products were enhanced with pictorial representations of alcohol content and health warnings, was carried out to inform this study (25). The review reported that a range of labelling approaches can effectively increase comprehension of the LRDG and the health risks of alcohol, particularly approaches that use pictorial warnings and messages relating to cancer. The authors concluded that the use of enhanced labels improves comprehension of unit information and the LRDG, especially when labels include information on both these things. It is possible that

including both of these components together in alcohol labels can enable a better understanding of units and of how many units one can consume within the LRDG. Further, although a growing body of research, including both quantitative and qualitative studies, suggests that adding health warnings to alcohol labels can increase perception of the health risks of alcohol consumption (21-24), no studies investigated effects on knowledge or understanding of LRDGs of adding health warnings to enhanced labels.

## Aims

The main aims of this trial were:

- 1) to compare the effectiveness of different label designs at conveying knowledge that the LRDG is 14 units
- 2) to compare the effectiveness of different label designs at conveying understanding of how many servings (bottle or can of beer, glass of wine, or shot of spirits) or containers (the entire bottle being purchased) could be consumed while remaining within the LRDG.

Secondary aims were:

- to compare the effect of the designs on the perceived risk of alcohol consumption
- to compare the effect of the designs on the motivation to drink
- to compare the effect of the designs on participants' perception of 'health-damaging' drinking (how many units per week they personally thought it would take for a person to 'seriously damage' their health).

Finally, we wanted to see whether showing people a health warning alongside our label designs would have a further effect on our secondary outcomes, increasing the perceived risk

of alcohol consumption, decreasing the motivation to drink, and lowering the level of drinking which people believe to be health-damaging. This was designed as a pilot study because we were not well powered; in particular we could not detect an interaction effect between the warning and the label designs, but we hoped to get some idea whether this hypothesis was worth pursuing in future trials.

## **2. Methods**

### *Study design*

This was a randomized controlled trial. When participants entered the survey, they were pseudorandomized using computerized random-number generation, which assigned them to one of seven arms (by assigning a number from 1 to 7), each of which saw a different label design. Once it had been determined which of the seven label-arms they would be in, a second random-number generation assigned some participants to also see a health warning underneath the design (participants were assigned a new random number between 1 and 100; those who got between 1 and 7 saw labels with the text, and those who got between 8 and 100 saw labels without the warning text). See the participant flow in Figure 1.

Participants did not know the nature of the other interventions. The task was described to them as a “survey” and they were not told what the other interventions were, or even that other participants might be seeing different labels. Immediately after participants saw the labels, in the same session, they were asked questions to determine their knowledge of the LRDG and their understanding of how much they could drink and stay under the LRDG.



There was an internal study protocol, which can be found in Appendix 1. The study was approved by the Research Support and Governance Office at Public Health England, Ref: R&D 347.

### *Participants*

Participants were recruited from 25 Jan - 1 Feb 2019. The trial ended when we had reached the number of responses determined by our power calculations. We recruited participants via a number of third-party panel providers, who have access to a pool of people who have given their consent to be contacted in order to answer online questionnaires. Participants were paid a fixed fee of approximately £1 for their time.

Participants were required to be English, over 18 years, and report drinking alcohol, as measured by the first question of the AUDIT-C questionnaire (26) (see procedure for full details of the screening). We specified that the sample should be representative of the adult population of England in terms of age, gender and region, which it was (see Table A1 in Appendix 2).

### *Interventions*

We compared the current industry standard and four other ways of showing information about alcohol content: pictograph, pie chart, risk gradient, and a design based on food labels. These were taken from current designs in the alcohol and food industries, other designs from the literature, and our bespoke pictograph designs. They were amongst a wider selection of designs that we showed to a focus group with 10 drinkers based in London in December 2018. (There were six males, four females; three 18-30 year olds, five 31-55 year olds and two over 55 year olds; representatives from all social classes A, B, C1, C2, D, E; seven White British,

three non-White ethnicity; three to four from each of low income <£25,000, middle income £25,000-£50,000, and high income > £50,000 groups; four low risk drinkers, 0-14 units per week, and six increased risk drinkers, 15-35 units for women, 15-50 units for men). We discarded the designs that the focus group participants considered too complicated and used their feedback to refine our preliminary pictograph designs.

We mainly showed the information in terms of servings (for each of the four designs and the control). However, since we were not certain whether showing the information by serving or container would be more effective, we also had one comparison of servings vs containers:

Pictograph serving and Pictograph container hold constant the way that the information is presented (pictograph style) but vary whether it is presented in terms of serving or container.

Further, since we wanted to know whether showing both pieces of information would be counterproductive, we had one comparison of single versus multiple framings of information:

Food label serving and Food label serving and container hold constant the way that the information is presented (food label style) but allow us to test the effect of giving participants only servings versus both serving and container information. This gives a total of seven different label designs for alcohol content, including the control.

Participants saw pictures of nine drinks, all seeing the same picture of the bottle and a box with information about the ABV and volume of the bottle. Alongside, they saw labels in one of the seven different label designs (see Figure 2 for examples):

1. Control (existing industry standard): outline of a bottle with the number of units that are in the entire bottle written inside the outline. No statement of the LRDG.
2. Food label serving: this design was based on food nutrition labels. There was a box that was split into two rows. On the top row was the number of units in a serving, on

the bottom row 'x% of the low risk drinking guidelines (14 units per week)'. Above the box, here was a picture of a serving (glass of wine, shot of spirits or bottle of beer) and information about the volume of a single serving.

3. Food label serving and container: As for food label serving, but now two boxes, one for servings and one for the container. On the left-hand side was a box showing the number of units in a serving, to exactly the same design as the Food label serving, including the picture above. To the right of this box, there was a similar box giving the same information for the container, i.e. the top row of the box had units per container, the bottom row had % of LRDG for the whole container (and repeated the information that the LRDG is 14 units per week), and above the box was a picture of the container and information about the volume of alcohol in the container.

4. Pictograph serving: this was a pictograph representation of the proportion of the LRDG that would be consumed in one serving. There was a picture of servings in outline (bottle/ can/ glass/ shots, as appropriate), with the first serving filled in black. The number of servings depicted varied, ranging from 5 to 26, so that, e.g., if one serving was  $\frac{1}{5}$  of the LRDG there would be 5 servings depicted with one filled in, or if one serving was  $\frac{1}{26}$  of the LRDG there would be 26 servings depicted with one of them filled in. Above the pictograph it said, '1 [serving] = [x] units'. The LRDG was written underneath the pictograph: 'The low risk drinking guideline is 14 units per week = [y servings]'. The number of servings in this phrase was the same as the number of outline servings in the pictograph.

5. Pictograph container: this was a pictograph representation of the proportion of the LRDG of the whole container's worth of beverage. There was a picture of containers in outline, filled in black to represent the proportion of a single container/ number of containers that would take one up to the LRDG. Above the pictograph it said: '1 bottle

= [x] units'. The LRDG was written underneath: 'The low risk drinking guideline is 14 units per week = [y] bottles'.

6. Pie chart serving: This was a pie chart that represented the proportion of LRDG in one serving. The number of slices in the pie varied, ranging from 5-26, the number being set so that one serving of the alcohol in question was one slice, so that e.g., if one serving was 1/5 of the LRDG the pie would be split into 5 slices with one filled in, or if one serving was 1/26 of the LRDG, the pie would be split into 26 slices with one of them filled in. The LRDG written underneath: 'The low risk drinking guideline is 14 units per week = [x servings]'. The number of servings in this phrase was the same as the number of slices in the pie.

7. Risk gradient serving: This had an x axis in the form of an arrow showing number of units, in colour, fading from yellow at just above zero, through orange, to red at 35, with 'low risk drinking guideline = 14 units per week' marked at 14 units, which was in the orange part of the spectrum. The number of units in a serving of the beverage was also marked on the axis. 'The more you drink, the greater the health risk' was written above the risk gradient axis.

For our pilot test, 500 participants (~70 in each condition) were randomly assigned to see one of the seven alcohol labels coupled with the text "Warning: Alcohol causes cancer" in bold, with a red line around it underneath the representation of alcohol content. (Figure 3 shows examples of how this appeared in the experiment).

### *Procedures*

Our experiment was conducted on the Behavioural Insight Team's online experimentation platform Predictiv.<sup>1</sup> The full materials are in Appendix 2.

Prior to the start of the survey, participants were screened using the first item of the Audit C questionnaire (27), "How often do you have a drink containing alcohol?". Anyone who answered "Never" was excluded from the survey, was not paid, and was not counted in the number of participants. Participants who passed the screening test were shown an information statement and asked if they consented to their data being used for research.

Participants were then randomized into one of seven conditions. The conditions were:

Control (existing industry standard), Food label serving, Food label serving and container, Pictograph serving, Pictograph container, Pie chart serving, Risk gradient serving. In addition, approximately 70 participants in each condition were randomized to also see a health warning underneath the label. Participants were shown nine pictures of drinks and their ABV, alongside an alcohol label; all nine labels used the design they had been allocated to (and the warning, if the participant had been allocated to that arm). There were 3 different beers, 3 different wines, and 3 different spirits; the drinks were the same for all participants, it was only the labels that changed. See Figure 2 for examples of the labels. The full set of labels is in Appendix 2. Participants could look at labels for as long as they liked and pressed "next" when they were ready to continue to the questions.

Then participants were asked about the knowledge primary outcome. After answering the knowledge question, participants were explicitly told that the LRDG was 14 units per week,

---

<sup>1</sup> Predictiv is an end-to-end platform that aims to make online experiments accessible to policy makers and other organisations driven by social impact. The platform provides functionality to run economic experiments and has access to a large international panel, including 200,000 people in the UK and 1 million in the US, through a network of online panel suppliers. More information can be found on [www.predictiv.co.uk](http://www.predictiv.co.uk).

before proceeding to ten understanding questions, which were presented in a random order.

Then participants were asked the secondary outcome questions, followed by some demographic questions. Finally, there was a free text box for feedback.

## *Measures*

### *Primary Outcomes*

#### 1. Knowledge of the LRDG

“The government’s low risk drinking guideline recommends that people not regularly drink more than a certain number of alcohol units per week. What do you think the low risk drinking guideline is?” (free text numeric response)

Our prespecified primary outcome measure for knowledge of the LRDG was whether participants gave the correct answer (binary variable, coded 1 if participant answered 14 units and 0 otherwise).

#### 2. Understanding of the LRDG

We asked ten understanding questions, which were presented in a random order. The general format of the questions was “How many [*serving/ container type (size in ml)*] of this [*beverage*] could you have before reaching 14 units?” (free text numeric response). We grouped the responses into two outcome measures, servings and containers. Note that we considered that a bottle/ can of beer was both a serving and a container, so the same two beer questions contributed to both the serving and the container measures.

##### 2a. Understanding (servings)

There were two questions on each of:

(i) beer: “How many bottles of this beer (330ml) could you have before reaching 14 units?”  
and “How many cans of this beer (586ml) could you have before reaching 14 units?”

(ii) wine: both “How many medium-sized glasses of this wine (175ml) could you have before reaching 14 units?”

(iii) spirits: both “How many single shots (25ml) of this drink could you have before reaching 14 units?”

For each of the six items we measured distance to the correct response by subtracting the answer given from the correct response (e.g., if the correct answer was 6 then a participant who entered 6 would get a score of zero, someone who entered 5 would get a score of -1, and someone who entered 10 would get a score of 4). Therefore, a positive score represents an overestimation and a negative score represents an underestimation. We then took an average of the six distances to calculate the outcome measure, which is a measure of number of servings from the correct answer.

We also decided to compare participants’ understanding of the LRDG measured in terms of units of alcohol, since for health purposes the number of units consumed is what matters. To do this, we converted the distance measure into units of alcohol, i.e. we calculated the number of units each participant was from the correct answer as expressed in units. Again, a positive score represents an overestimation and a negative score represents an underestimation. The score is a measure of the number of units from the correct answer.

## 2a. Understanding (containers)

There were two questions on each of:

(i) beer: “How many bottles of this beer (330ml) could you have before reaching 14 units?” and “How many cans of this beer (586ml) could you have before reaching 14 units?”

(ii) wine: both “How many bottles of this wine (750ml) could you have before reaching 14 units?”

(iii) spirits: “How much of a bottle or whole bottles (700ml) could you have before reaching 14 units?” and “How much of a bottle or whole bottles (1L) could you have before reaching 14 units?”

For each of the six items we measured distance to the correct response by subtracting the answer given from the correct response, as detailed for the servings measure 2a, and took an average of the six distances to calculate the outcome measure, measured in number of containers from the correct answer. We also converted the distance measure for containers into units of alcohol, to get the score in terms of the number of units of alcohol from the correct answer, as for the servings measure 2a.

### *Secondary outcomes*

Our secondary outcomes were:

(i) *perceived personal risk*

“To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease?”

Scale of 1 = Not at all likely, 2 = Not very likely, 3 = Somewhat likely, 4 = Quite likely, 5 = Extremely likely.

(ii) *motivation to drink*

“Earlier, you saw the following alcohol label: [beer image #3].



To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less.”

Scale of 1 = Strongly disagree 2 = Disagree 3 = Neither agree nor disagree 4 = Agree 5 = Strongly agree.

(iii) *perception of “damaging” drinking*

“How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health?” (free text numeric response)

*Demographics*

Participants completed the full Audit C questionnaire, provided demographic information on profession/social grade, smoking status (not presented), ethnicity, highest level of educational attainment (the recruitment companies already had age, gender, and which region of the UK the participant lives in). There was an attention check question amongst the demographic items.

Finally, participants were asked for any feedback about the label, in an open-text box, for instance whether they found it useful or confusing, or whether they thought it should be changed.

**Statistical Analysis**

*Sample size*

A pre-trial power calculation showed that 1000 participants in each arm was sufficient to identify an increase of 4.5%-6.4% in the participants who correctly identified the LRDG as 14

units per week, with 80% power and an alpha level of between 0.2% and 5% (we adjusted alpha to account for multiple comparisons, using a Hochberg step-up procedure), assuming that 13% of participants in the baseline condition, who saw the existing industry-standard labels, would correctly identify the LRDG as 14 units per week. We also recruited a further 500 participants (approximately 70 in each condition) for a pilot investigation, which included a warning about health risks alongside the label.

### *Statistical Analysis*

In order to test knowledge of the LRDG, we ran a logistic regression with whether or not the participant gave the correct answer as the dependent variable, controlling for demographic characteristics, AUDIT-C, and warning labels. In order to test understanding, we ran an OLS regression for each of our distance measures (servings and containers), controlling for demographic characteristics, AUDIT-C, and warning labels. For our secondary measures, we ran OLS regressions, controlling for demographic characteristics, AUDIT-C, and warning labels. Data were analyzed in Stata 14.2. The analysis plan was prespecified in an internal trial protocol (Appendix 1), but it was not pre-registered on a publicly available platform, so the results could be considered exploratory. Post hoc, we ran exploratory OLS regressions of our understanding measures disaggregated into different types of alcohol and also with the measures converted into number of units. On the request of reviewers we added a comparison of the proportion in each condition who over- vs under-estimated the LRDG, given that they had got the answer wrong.

## **3. Results**

### *Participants*

We analysed the data of 7516 participants. We excluded 504 participants because they failed the attention check (6.3% of the total 8025 who completed the survey). We excluded a further five participants because their free text numeric response answers were outliers and their survey responses suggested that they had not made a serious attempt to answer the questions (for more detail see Appendix 1). The participant flow is shown in Figure 1.

Participants were recruited via a number of panel providers. Eligible participants were English, over 18 years, and drinkers of alcohol. There were 3798 women and participants were aged between 18-99 years ( $M = 44.15$ ,  $SD = 16.45$ ). Details of our participants' baseline characteristics can be found in Table 1. Our sample was recruited to be representative of the adult population of England in terms of age, gender and region.

*Primary outcome: Knowledge of LRDG*

More participants underestimated than overestimated the LRDG and the distribution was skewed (see Figure 5): the modal response was the correct answer of 14, the median was 12 and the interquartile range was 9 (from 5-14).

In the control group, only 21.3% of participants correctly answered that the LRDG was 14 units per week. A logistic regression showed that participants in all of the intervention conditions had a more accurate knowledge of the LRDG than those in the control condition (all  $p < 0.001$ , summary statistics and Adjusted Odds Ratios are reported in Table 2). There appears to be a cluster of three best-performing designs (Pictograph container, 51.1%, followed by Pictograph serving 48.8%, and Pie chart serving, 47.5%—the three that had the LRDG in a separate statement, underneath the graphics), and three that did not perform quite so well, even though they performed better than the control (Food label serving 38.7%, Risk

gradient serving, 35.6%, and Food label serving and container, 32.9%), as shown in Figure 4, where the unadjusted 95% confidence intervals do not overlap between the two clusters or the control.

Over 80% of those who gave the wrong LRDG gave an answer that was less than the LRDG; this was true in all conditions (see Figure 6). Although the proportion who got the LRDG correct varied depending on the label design, given that participants had got the answer wrong, there were no statistically significant differences in whether they were likely to under- or over-estimate between conditions,  $\chi^2(6) = 10.22, p = 0.11$ .

Several of the variables that we controlled for in the OLS regression were related to knowledge. Those who were 55+ were more likely to answer the question correctly than 18 to 24-year olds, and people with any level of education from secondary upwards were more accurate than people with no secondary education. Lower social grades (C2DE) answered less accurately than higher grades (ABC1); Black, Asian, and Mixed-race ethnicities less accurately than White. There were regional variations. There was no statistically significant relationship between answer to the knowledge measure and sex, Audit-C score, or having seen the warning.

#### *Primary outcome: Understanding of LRDG servings*

The understanding (servings) measure had good internal consistency (Cronbach's alpha = 0.67).

Every group underestimated how many servings it takes to reach 14 units (see Table 3).

Control group participants were the least accurate on our primary outcome measure (the

average of their distance measures for two beers, two wines, and two spirits), they estimated they could have  $M = 4.64$ ,  $SD = 3.43$  fewer servings than the LRDG 14 units ( $M = 4.43$  fewer units,  $SD = 3.95$ ). Participants in the best performing group, Pictograph serving, thought they could have  $M = 0.93$ ,  $SD = 3.43$  fewer servings than the LRDG 14 units ( $M = 0.96$  fewer units,  $SD = 2.46$ ). An OLS regression showed that participants in all of the interventions had a better understanding of how many servings they could consume and remain under the 14-unit LRDG than those in the control condition (all  $p < 0.001$ , see Table 4 for full model and confidence intervals). Comparing the four intervention designs that only gave information in terms of servings, Risk gradient serving performed the worst—it did not have overlapping confidence intervals with any of the other three for accuracy of number of servings in the adjusted model—and the numerical ordering of performance was Pictograph serving > Pie chart serving > Food label serving > Risk gradient serving. There was no evidence of any detriment in understanding of LRDG servings from adding container information to the food label design: Food label serving had overlapping confidence intervals with Food label serving and container in the adjusted model). It is notable that the Pictograph container condition, the only intervention not to give information in servings, while more accurate than the control, was less accurate than all the other intervention arms (no overlapping 95% confidence intervals, either adjusted or unadjusted). There was no effect of having seen a warning label. The inaccuracy was driven by the estimates for wines and especially spirits. Figure 7 shows the understanding estimates for servings, disaggregated into wine, beer, and spirits. For beer, all intervention groups gave similar and accurate answers to questions about how many servings they could have. Within each label design the confidence intervals of the understanding estimates for servings of beer, alcohol and spirits do not overlap, with participants being least accurate about servings of spirits. When the estimates are expressed

in terms of units, the numerical ordering is preserved, but some of the confidence intervals overlap (see Figure 7).

*Primary outcome: Understanding of LRDG containers*

The understanding (containers) measure had good internal consistency (Cronbach's alpha = 0.66).

Every group overestimated how many containers it takes to reach 14 units (see Table 3).

Control participants were the most accurate, when averaging across their estimates for how many containers of beer, wine, and spirits they could have, they estimated they could have  $M = 0.09$ ,  $SD = 1.02$  more containers ( $M = 6.00$  more units,  $SD = 14.08$ ) than actually allowed.

Participants in the numerically worst performing group, Food label serving, thought they could have  $M = 1.10$ ,  $SD = 1.27$  too many containers ( $M = 19.62$  fewer units,  $SD = 20.36$ ). An OLS regression showed that participants in all of the interventions had a worse understanding of how many containers they could consume and remain under the LRDG than those in the control condition (all  $p < 0.001$ , see Table 4 for the full model including confidence intervals).

The most accurate two intervention conditions were Pictograph container and Food labels servings and container, the two that gave information in terms of containers, which had 95% confidence intervals that did not overlap with any other of the other interventions (though the adjusted confidence intervals overlapped with each other). Comparing the four designs that only gave information in terms of servings, the numerical ordering of performance is Pie chart serving > Risk gradient serving > Pictograph serving > Food label serving, though there were overlaps in the 95% confidence intervals of the coefficients in the adjusted model. There was no effect of having seen a warning label.

Again, these inaccuracies were driven almost entirely by participants' estimates for wine and spirits. The beer estimates were most accurate in all conditions and, within each condition, the confidence intervals for beer estimates did not overlap with those for wine or spirits (See Figure 8). When estimates were expressed in terms of containers, the wine estimates were numerically most inaccurate and confidence intervals did not overlap with spirits estimates in any condition except the Control and Pictograph containers. When estimates were expressed in terms of units, then spirits estimates were numerically most inaccurate and the confidence intervals did not overlap with wine for any condition apart from Food servings and containers.

### *Secondary outcomes*

For our secondary measures, we found that participants in all conditions on average thought it was "quite likely" that cutting down on their alcohol consumption would reduce the risk of disease ( $M = 3.88$ ,  $SD = 1.22$ ), they on average neither agreed nor disagreed that the alcohol label made them less motivated to drink ( $M = 3.23$ ,  $SD = 1.03$ ), and the average estimate of how many units per week a person would need to drink to seriously damage their health was 24 units ( $M = 26.24$ ,  $SD = 62.60$ ). (See Table 6 for a complete breakdown by trial arm.) We ran OLS regressions on the secondary measures and found that the enhanced label designs had no effect on the perceived personal risk of drinking or on the perception of health-damaging drinking, but they all decreased stated motivation to drink compared to the control, albeit by a very small amount (0.1 - 0.3 points on a 5-point scale). There was no effect of the warnings. For the full models see Table 6.

### **Discussion**

All of our enhanced alcohol-label designs improved knowledge of the LRDG. In the Control group, only 21.5% of participants correctly answered that the LRDG was 14

units per week, but knowledge was higher in every intervention arm (proportion of correct answers ranged from 32.9% to 51.5%). Our enhanced designs improved understanding of the LRDG when that was expressed in terms of servings but decreased understanding when it was expressed in terms of containers. The enhanced designs had no effect on the perceived personal risk of drinking or on the subjective perception of high-risk drinking, but they all decreased stated motivation to drink compared to the control, albeit by a very small amount. The addition of a cancer warning had no effect on any of our measures.

It is not surprising that our interventions increased the level of knowledge of the LRDG, since the existing industry-standard label was the only one that did not explicitly state the LRDG. The 21.5% who responded with the correct LRDG in the control condition is consistent with the results of recent UK surveys, where the proportion of participants correctly reporting the LRDG has varied from 8% to 25% (16-18). Of our new enhanced designs, Pictograph servings, Pictograph container, and Pie chart fared particularly well, with 47-51% of participants correctly reporting the LRDG. In all three of these designs, the LRDG was given in a separate statement, underneath the graphics, which may have made it particularly salient. Participants who gave an incorrect answer were more likely to underestimate than overestimate the LRDG—in all conditions, even as the number giving an accurate answer increased, more than 80% of those who were incorrect gave an underestimate.

Participants in all seven conditions underestimated the number of servings that they could drink and still remain under the LRDG. Understanding in terms of servings was more accurate in the intervention groups. This replicates the findings of two



previous studies (28, 29). Conversely, participants in all conditions overestimated the number of containers of alcohol that they could drink and still remain under the LRDG, and accuracy of understanding decreased in the intervention conditions. When we disaggregated in terms of type of alcohol, we found that participants' estimates for beer were similar and fairly accurate across conditions; the inaccuracies and differences were driven by estimates for wine and especially spirits (when denominated in terms of number of units). One reason why our participants were more accurate for beer may be because people often drink entire containers (bottles or cans) of beer as a single serving, but they usually need to pour out servings of wine and spirits, and they rarely drink a whole container of spirits in one sitting. This suggests that, potentially, we could improve understanding of the alcohol content of wine and spirits if containers and serving vessels had lines indicating standard units, so that people are more aware of the number of servings they are consuming.

The most effective labels differed depending on whether understanding was measured in terms of units or of containers. Unsurprisingly, the accuracy of understanding estimates varied depending on whether the design participants had seen was congruent with the question: the designs showing servings led to more accurate answers to questions about servings, whilst designs showing containers led to more accurate estimates of containers. Pictographs were highly successful when the presentation and the question were congruent but amongst the least successful labels when they were not congruent. Pie chart servings was a reasonable performer on both understanding measures. Interestingly, adding container information to the food label design, as well as servings, increased understanding of containers without any detriment in understanding of servings. However, we cannot infer that providing both types of

information on the other designs, which had more reliance on graphics, would have the same effect, though this might be worthy of further investigation.

From a public health perspective, when deciding whether to present information in terms of servings or containers, it would be better to choose whichever keeps health risks lower. There are two considerations, which pull in opposite directions. In order to encourage lower alcohol consumption, it is better if participants underestimate (rather than overestimate) how much they can drink, which suggests contextualising how much alcohol it takes to reach the LRDG in terms of servings. However, total alcohol intake will depend not only on how many servings people think they can have, but also on whether they can accurately track how many servings they are drinking. People tend to have difficulty pouring standard drinks, with over-pouring being the norm (20). Recent studies have found that when people are asked to pour out a 'normal serving' of spirits, they on average pour 2 units of alcohol, similar to a 50ml double shot (30, 31). Even if people underestimate the amount of servings they are allowed, if they overpour their drinks (overestimating the size of a standard serving), then labels that are denominated in terms of servings may lead to higher alcohol consumption than labels that are denominated in terms of containers.

The health warning did not affect responses to either the primary or secondary outcome measures. It seems likely that this was despite participants noticing it, since the warning was large and in a red box, and both size and colour have been shown to be important in whether people pay attention to warnings (32, 33). Although other studies have found that cancer warnings increase the perceived risk of drinking alcohol (21), reduce stated motivation to drink (28), and reduce stated future drinking intentions (23), in those studies the cancer warnings were always being compared to other types of warnings. Two other studies

presented participants with warning labels compared to a control condition with no label and both found no effect of the text warning compared to the control (24, 34). Pictures and graphical warnings have been found to be more effective than text warnings (24, 35). Further, it may not be surprising that a warning alone has no effect, since there is a large body of evidence on “fear appeals”, which shows that fear-control processes can interfere with the motivation to take precautions (36) and that fear appeals are only effective in the presence of high self-efficacy for taking action to prevent the risk (37, 38); where there is low self-efficacy, fear appeals may lead to avoidance or reactance (37).

We found that enhanced labels can improve knowledge and understanding of the LRDG, but they did not affect our secondary outcomes: the perceived risk of alcohol consumption, the motivation to drink, and the level of drinking which people believe to be health-damaging. In general, our participants tended to underestimate both the LRDG and of the number of servings they can drink and remain beneath it, which may be protective. Although our results suggest that improving alcohol labelling alone is unlikely to change behaviour, enhanced labels could still facilitate informed choice. This raises the prospect that improving knowledge and understanding might lead to an increase in alcohol consumption, if people adjust consumption upwards to reach the LRDG. Therefore, it is important that people understand the nature of the dose-response relationship, whereby risk increases with drinking, rather than regarding the LRDG as a threshold for safe drinking.

We randomized a large number of individuals to each condition, which is a strength of our trial. The main limitation of our trial is that we ran an online experiment and our results may not generalise well to field settings, including supermarkets, which are the places that people are most likely to see a label of the sort we tested, on a bottle before they buy it. In our online setting, the labels were presented on a screen, and although the size of the labels on mobile

screens were comparable to the size of labels on a bottle, participants could zoom in if they wished. Labels that display a large amount of information (e.g. the Food label serving and container design) or which are wide in design (e.g. the Risk gradient) may perform worse in real life if they need to be shrunk down to fit on standard alcohol packaging. Furthermore, although the average participant in our experiment spent around 60 seconds reviewing the various example labels, we know from laboratory studies that people do not pay much attention to alcohol health warnings or responsible drinking statements, (32, 39) and we expect that they would pay even less attention to them at point of sale. Lastly, although our sample was designed to be representative of basic population characteristics, our participants were a self-selecting group who had agreed to be on a panel and answer questions for money. Potentially their behaviour may not be representative of the average member of the population. So, although our study shows that our labels would improve knowledge and understanding if people pay attention to them, we cannot be sure to what extent those results would generalise to a field setting where people might not pay attention.

This study was about comprehension of risk. Although we asked about intention to reduce alcohol consumption, not only did we not find a meaningful effect of the labels, but we also know that there is an intention-behaviour gap: stated intentions may not translate into behaviour (40). As well as testing comprehension at point of sale and investigating how to get people to pay more attention to labels in the field, future research could investigate whether different label designs have any effect on purchasing and consumption behaviour.

Taken together, these results show that improved pictorial designs to communicate alcohol risk can lead to better knowledge and understanding of LRDGs. All of our custom designs improved knowledge that the UK LRDG is 14 units, compared to industry-standard labels.

Designs that had the LRDG in a separate statement, underneath the graphics, improved knowledge the most. For understanding, different designs performed best depending whether the question was how many servings could be consumed while remaining under the LRDG or how many containers (and the safe number of servings was underestimated, so improving understanding could increase the amount that people think it is safe to drink). However, the results suggest there is room for improvement in existing alcohol labels.

### **Acknowledgments**

We thank the Bristol Tobacco and Alcohol Research Group for input on the label designs, Lucy Porter for designing the CONSORT flowchart, Adam Winter on commissioning and conception of the design and Jeric Kison for comments on the paper and supporting the project management. FG's research is supported by the National Institute for Health Research Applied Research Collaboration Northwest London. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

### **Funding**

Public Health England.

### **References**

1. World Health Organisation. Global Status Report on Alcohol and Health. 2018.
2. Babor T, Caetano R, Casswell S, Edwards G, Giesbrecht N, Graham K, et al. Alcohol: No ordinary commodity: Research and public policy New York: Oxford: Oxford University Press; 2010.
3. Rehm J, Taylor B, Mohapatra S, Irving H, Baliunas D, Patra J, et al. Alcohol as a risk factor for liver cirrhosis: a systematic review and meta-analysis. *Drug alcohol review*. 2010;29(4):437-45.
4. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, et al. Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. *British Journal of Cancer*. 2015;112(3):580.
5. Burton R, Henn C, Lavoie D, O'Connor R, Perkins C, Sweeney K, et al. A rapid evidence review of the effectiveness and cost-effectiveness of alcohol control policies: an English perspective. *The Lancet*. 2017;389(10078):1558-80.
6. Furtwängler NA, de Visser RO. Lack of international consensus in low-risk drinking guidelines. *Drug & Alcohol Review*. 2013;32(1):11-8.
7. Kalinowski A, Humphreys K. Governmental standard drink definitions and low-risk alcohol consumption guidelines in 37 countries. *Addiction*. 2016;111(7):1293-8.

8. Babor T, Higgins-Biddle J. Brief intervention for hazardous and harmful drinking: a manual for use in primary care; 2001. Geneva: World Health Organization. 2011.
9. Lord President of the Council. Action Against Alcohol Misuse. London: HMSO; 1991.
10. Department of Health. The Health of the Nation - A strategy for health in England. London: HMSO; 1992.
11. Department of Health. UK Chief Medical Officers' Low Risk Drinking Guidelines. Department of Health London; 2016.
12. Health and Social Care Information Centre. Health Survey for England 2014: Chapter 8 Adult Alcohol Consumption. 2015.
13. Department of Health. Communicating the UK Chief Medical Officers' low risk drinking guidelines. 2017.
14. Department of Health. Public Health Responsibility Deal: Alcohol Pledges. . 2011.
15. Burton R, Henn C, Lavoie D, O'Connor R, Perkins C, Sweeney K, et al. A rapid evidence review of the effectiveness and cost-effectiveness of alcohol control policies: an English perspective. *The Lancet*. 2017;389(10078):1558-80.
16. Rosenberg G, Bauld L, Hooper L, Buykx P, Holmes J, Vohra J. New national alcohol guidelines in the UK: public awareness, understanding and behavioural intentions. *Journal of Public Health*. 2017;40(3):549-56.
17. Alcohol Health Alliance UK. How we drink, what we think. Public views on alcohol and alcohol policies in the UK. 2018.
18. Buykx P, Li J, Gavens L, Hooper L, Gomes de Matos E, Holmes J. Self-reported knowledge, correct knowledge and use of UK drinking guidelines among a representative sample of the English population. *Alcohol and Alcoholism*. 2018;53(4):453-60.
19. Department of Health. Alcohol Units: a brief guide. 2008.
20. Kerr WC, Stockwell T. Understanding standard drinks and drinking guidelines. *Drug & Alcohol Review*. 2012;31(2):200-5.
21. Jongenelis MI, Pratt IS, Slevin T, Chikritzhs T, Liang W, Pettigrew S. The effect of chronic disease warning statements on alcohol-related health beliefs and consumption intentions among at-risk drinkers. *Health Education Research*. 2018;33(5):351-60.
22. Miller ER, Ramsey IJ, Baratin GY, Olver INJBph. Message on a bottle: are alcohol warning labels about cancer appropriate? 2016;16(1):139.
23. Pettigrew S, Jongenelis M, Chikritzhs T, Slevin T, Pratt IS, Glance D, et al. Developing cancer warning statements for alcoholic beverages. *BMC Public Health*. 2014;14(1):786.
24. Wigg S, Stafford LD. Health warnings on alcoholic beverages: perceptions of the health risks and intentions towards alcohol consumption. *PLOS One*. 2016;11(4):e0153027.
25. Burton R, Smolar M, Gold N, Harper H, Kroner Dale M, Brown H, et al. The effectiveness of alcohol label information, warnings and risk communication: a rapid evidence review. under review.
26. Bush K, Kivlahan DR, McDonnell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Archives of internal medicine*. 1998;158(16):1789-95.
27. [Available from:  
[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/684826/Alcohol\\_use\\_disorders\\_identification\\_test\\_for\\_consumption\\_AUDIT\\_C.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/684826/Alcohol_use_disorders_identification_test_for_consumption_AUDIT_C.pdf).

28. Blackwell AK, Drax K, Attwood AS, Munafò MR, Maynard OM. Informing drinkers: Can current UK alcohol labels be improved? *Drug and Alcohol Dependence*. 2018;192:163-70.
29. Hobin E, Vallance K, Zuo F, Stockwell T, Rosella L, Simniceanu A, et al. Testing the efficacy of alcohol labels with standard drink information and national drinking guidelines on consumers' ability to estimate alcohol consumption. *Alcohol and Alcoholism*. 2017;53(1):3-11.
30. Gill JS, Donaghy M. Variation in the alcohol content of a 'drink' of wine and spirit poured by a sample of the Scottish population. *Health Education Research*. 2004;19(5):485-91.
31. Boniface S, Kneale J, Shelton N. Actual and Perceived Units of Alcohol in a Self-Defined "Usual Glass" of Alcoholic Drinks in England. *Alcoholism: Clinical and Experimental Research*. 2013;37(6):978-83.
32. Pham C, Rundle-Thiele S, Parkinson J, Li S. Alcohol warning label awareness and attention: a multi-method study. *Alcohol and Alcoholism*. 2017;53(1):39-45.
33. Al-Hamdani M, Smith SM. Alcohol warning label perceptions: do warning sizes and plain packaging matter? *Journal of studies on alcohol and drugs*. 2016;78(1):79-87.
34. Krischler M, Glock S. Alcohol warning labels formulated as questions change alcohol-related outcome expectancies: A pilot study. *Addiction Research & Theory*. 2015;23(4):343-9.
35. Chen Y, Yang ZJ. Message formats, numeracy, risk perceptions of alcohol-attributable cancer, and intentions for binge drinking among college students. *Journal of Drug Education*. 2015;45(1):37-55.
36. Ruiter RA, Abraham C, Kok G. Scary warnings and rational precautions: A review of the psychology of fear appeals. *Psychology & Health*. 2001;16(6):613-30.
37. Witte K, Allen M. A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health education & behavior*. 2000;27(5):591-615.
38. Tannenbaum MB, Hepler J, Zimmerman RS, Saul L, Jacobs S, Wilson K, et al. Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*. 2015;141(6):1178.
39. Kersbergen I, Field M. Visual attention to alcohol cues and responsible drinking statements within alcohol advertisements and public health campaigns: Relationships with drinking intentions and alcohol consumption in the laboratory. *Psychology of Addictive Behaviours*. 2017;31(4):435.
40. Sheeran P, Webb TL. The intention-behavior gap. *Social and personality Psychology*. 2016;10(9):503-18.

Table 1: Baseline demographics characteristics of the seven trial arms and overall for the whole trial

<b>Trial arm</b>	<b>Number in trial arm</b>	<b>Number (%) of females</b>	<b>Age Mean (SD)</b>	<b>Audit-C score Mean (SD)</b>
Control	1044	516 (50.6%)	44.18 (16.75)	4.96 (2.67)
Food label (servings)	1074	558 (52.0%)	43.58 (16.23)	5.04 (2.69)
Food label (servings and containers)	1120	569 (50.8%)	44.05 (16.34)	5.00 (2.70)
Pictograph (containers)	1085	571 (52.6%)	43.94 (16.35)	5.09 (2.71)
Pictograph (servings)	1089	543 (49.9%)	43.94 (16.56)	5.17 (2.75)
Pie chart (servings)	1062	525 (49.4%)	44.15 (16.31)	5.09 (2.78)
Risk gradient (servings)	1042	516 (49.5%)	45.26 (16.61)	5.03 (2.73)
<b>Overall</b>	<b>7516</b>	<b>3798 (50.5%)</b>	<b>44.15 (16.45)</b>	<b>5.06 (2.72)</b>



Table 2: Knowledge of the LRDG: Proportion of participants who correctly identified the LRDG as 14 units and Adjusted Odds Ratios from a binary logistic regression controlling for demographics; ordered from smallest to largest AOR

<b>Trial arm</b>	<b>Number of participants in the trial arm</b>	<b>Number correctly identifying LRDG</b>	<b>% correctly identifying LRDG</b>	<b>Adjusted Odds Ratio</b>	<b>95% CIs</b>		<b>p-value</b>
Control	1044	222	21.3	---	---	---	---
Food label Servings and Containers	1120	368	32.9	1.85	1.52	2.26	< 0.001
Risk Gradient	1042	371	35.6	2.09	1.71	2.55	< 0.001
Food Label	1074	416	38.7	2.44	2.01	2.97	< 0.001
Serving	1062	504	47.5	3.57	2.93	4.34	< 0.001
Pie Chart	1089	531	48.8	4.11	3.39	4.99	< 0.001
Serving	1085	554	51.1	3.74	3.08	4.54	< 0.001
Pictograph							
Container							

Table 3: Understanding of the LRDG: Distance from the correct answer for questions about how many servings /containers could be consumed before reaching 14 units (each measure is an average of six answers: two beer, two wine, and two spirits)

Trial arm (ordered most to least accurate)	Accuracy of understanding (servings)					
	Servings			Units		
	Mean (SD)	95% CIs		Mean (SD)	95% CIs	
Pictograph serving	-0.93 (2.17)	-1.06	-0.80	-0.96 (2.46)	-1.10	-0.81
Pie chart serving	-1.11 (2.49)	-1.26	-0.96	-1.12 (2.93)	-1.30	-0.94
Food label serving	-1.21 (2.75)	-1.37	-1.04	-1.20 (3.02)	-1.38	-1.02
Food label serving and container	-1.40 (2.85)	-1.56	-1.23	-1.36 (3.10)	-1.54	-1.18
Risk gradient serving	-1.84 (3.63)	-2.06	-1.62	-1.61 (4.91)	-1.91	-1.31
Pictograph container	-3.45 (3.44)	-3.66	-3.25	-2.96 (4.20)	-3.21	-2.71
Control	-4.64 (3.43)	-4.85	-4.44	-4.43 (3.95)	-4.67	-4.19

Trial arm (ordered most to least accurate)	Accuracy of understanding (containers)					
	Containers			Units		
	Mean (SD)	95% CIs		Mean (SD)	95% CIs	
Control	0.09 (1.02)	0.03	0.16	6.00 (14.08)	5.14	6.85
Pictograph container	0.22 (0.99)	0.16	0.27	6.44 (15.21)	5.54	7.35
Food label serving and container	0.40 (1.09)	0.33	0.46	8.31 (15.44)	7.41	9.22
Pie chart serving	0.80 (1.17)	0.73	0.87	14.81 (18.47)	13.70	15.92
Risk gradient serving	0.81 (1.56)	0.72	0.91	15.74 (20.14)	14.51	16.96
Pictograph serving	0.90 (1.13)	0.84	0.97	15.78 (18.63)	14.68	16.89
Food label serving	1.10 (1.27)	1.02	1.17	19.62 (20.36)	18.40	20.84

Table 4. Understanding of the LRDG: OLS regression with accuracy of estimate of how many servings/ containers could be drunk and the drinker still remain under the 14 unit per week LRDG

Characteristic	Servings: Distance to correct answer <sup>1</sup> (compared to baseline category for categorical variables)				Containers: Distance to correct answer <sup>2</sup> (compared to baseline category for categorical variables)			
	$\beta$ (SE)	95% CIs		p-value	$\beta$ (SE)	95% CIs		p-value
<b>Treatment (baseline = Control)</b>								
Food label serving	3.42 (0.13)	3.16	3.67	<0.001	1.02 (0.05)	0.92	1.12	<0.001
Food label serving and container	3.24 (0.13)	2.99	3.49	<0.001	0.32 (0.05)	0.22	0.42	<0.001
Pictograph serving	3.70 (0.13)	3.44	3.95	<0.001	0.82 (0.05)	0.72	0.92	<0.001
Pictograph container	1.17 (0.13)	0.92	1.43	<0.001	0.14 (0.05)	0.04	0.24	0.007
Pie chart serving	3.53 (0.13)	3.27	3.78	<0.001	0.72 (0.05)	0.62	0.82	<0.001
Risk gradient serving	2.79 (0.13)	2.54	3.05	<0.001	0.74 (0.05)	0.64	0.84	<0.001
<b>Age (baseline = 18-24)</b>								
25-54	-0.09 (0.11)	-0.29	0.12	0.42	-0.30 (0.4)	-0.38	-0.21	<0.001
55+	0.42 (0.11)	0.20	0.64	<0.001	-0.33 (0.5)	-0.42	-0.24	<0.001
<b>Female (baseline = male)</b>	0.2 (0.7)	-0.12	0.16	0.75	-0.09 (0.03)	-0.15	-0.04	0.001
<b>Social grade C2DE (baseline = ABC1)</b>	-0.27 (0.07)	-0.41	-0.12	<0.001	-0.02 (0.03)	-0.08	0.04	0.47
<b>Ethnicity (baseline = White)</b>								
Black	-0.62 (0.23)	-1.07	-0.18	0.006	0.44 (0.09)	0.26	0.62	<0.001
Asian	-0.57 (0.19)	-0.94	-0.21	0.002	0.34 (0.07)	0.19	0.48	<0.001
Mixed	-0.65 (0.24)	-1.12	-0.17	0.007	0.23 (0.10)	0.04	0.41	0.018
Other	-0.11 (0.40)	-0.90	0.68	0.78	0.35 (0.16)	0.04	0.67	0.027
<b>Region (baseline = North)</b>								
South & East	0.22 (0.09)	0.05	0.38	0.012	-0.03 (0.03)	-0.10	0.04	0.35
Midlands	-0.15 (0.10)	-0.35	0.04	0.12	-0.07 (0.04)	-0.14	0.01	0.097
London	-0.40 (0.11)	-0.61	-0.16	0.001	0.05 (0.05)	-0.03	0.14	0.23
<b>Audit C (numerical, 1-12)</b>	0.02 (0.01)	-0.01	0.04	0.20	0.00 (0.01)	-0.01	0.01	0.997
<b>Highest education (baseline = none)</b>								
Secondary	0.67 (0.26)	0.16	1.18	0.01	-0.05 (0.10)	-0.25	0.15	0.62
Post-secondary / Vocational	1.21 (0.26)	0.70	1.71	<0.001	-0.12 (0.10)	-0.32	0.08	0.26
Undergrad or higher	1.67 (0.26)	0.76	1.78	<0.001	-0.21 (0.10)	-0.42	-0.01	0.038

<b>Warning (baseline = no warning)</b>	0.05 (0.14)	-0.23	0.32	0.75	0.04 (0.05)	-0.06	0.15	0.42
<b>Constant</b>	-5.73 (0.30)	-6.33	-5.14	<0.001	0.52 (0.12)	0.29	0.76	<0.001
<b>R-squared</b>		0.18				0.10		
<b>Sample size</b>		<b>7481</b>				<b>7500</b>		

<sup>1</sup>The ‘Servings’ outcome was measured by taking the average of people’s estimates for how many beers (2 questions), servings of wines (2 questions), and servings of spirits (2 questions) it takes to reach 14 units, and then subtracting the technically correct answer from this. The analysis excludes 35 participants who gave ineligible responses for at least one of these 6 questions.

<sup>2</sup>The ‘Containers’ outcome was measured by taking the average of people’s estimates for how many beers (2 questions), containers of wines (2 questions), and containers of spirits (2 questions) it takes to reach 14 units, and then subtracting the technically correct answer from this. The analysis excludes 16 participants who gave ineligible responses for at least one of these 6 questions.

Table 5: Secondary outcomes (OLS regressions): Perceived personal risk of own drinking (1-5), motivation to drink (1-5), and subjective perception of high-risk drinking (numeric free text response)

Trial arm	Perceived risk <sup>1</sup>			Motivation to drink <sup>2</sup>			Perception of health-damaging drinking <sup>3</sup>		
	Mean (SD)	95% CIs		Mean (SD)	95% CIs		Mean (SD)	95% CIs	
Control	3.87 (1.16)	3.80	3.94	3.07 (1.08)	3.00	3.13	25.00 (36.50)	22.78	27.22
Food label serving	3.84 (1.13)	3.77	3.90	3.21 (1.04)	3.14	3.27	26.02 (46.93)	23.21	28.83
Food label serving and container	3.89 (1.11)	3.83	3.96	3.23 (1.03)	3.17	3.29	24.88 (23.51)	23.51	26.26
Pictograph serving	3.89 (1.10)	3.83	3.96	3.23 (1.04)	3.16	3.29	25.30 (21.02)	24.05	26.55
Pictograph container	3.87 (1.11)	3.80	3.93	3.33 (1.00)	3.27	3.39	26.22 (48.91)	23.30	29.13
Pie chart serving	3.90 (1.11)	3.83	3.96	3.29 (0.99)	3.23	3.35	26.03 (25.69)	24.48	27.57
Risk gradient serving	3.91 (1.12)	3.85	3.98	3.27 (1.05)	3.20	3.33	23.90 (17.11)	22.86	24.94
Overall average	3.88 (1.12)	3.86	3.91	3.23 (1.03)	3.21	3.26	25.34 (33.54)	24.58	26.10

<sup>1</sup>To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease? From 1 (Not at all likely) to 5 (Extremely likely)

<sup>2</sup>Earlier, you saw the following alcohol label: [beer image #3]. To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less. From 1 (Strongly disagree) to 5 (Strongly agree)

<sup>3</sup>How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health? Free text numeric response

Table 6. Secondary outcomes (OLS regressions): Perceived personal risk of own drinking (1-5), motivation to drink (1-5), and subjective perception of high-risk drinking (numeric free text response)

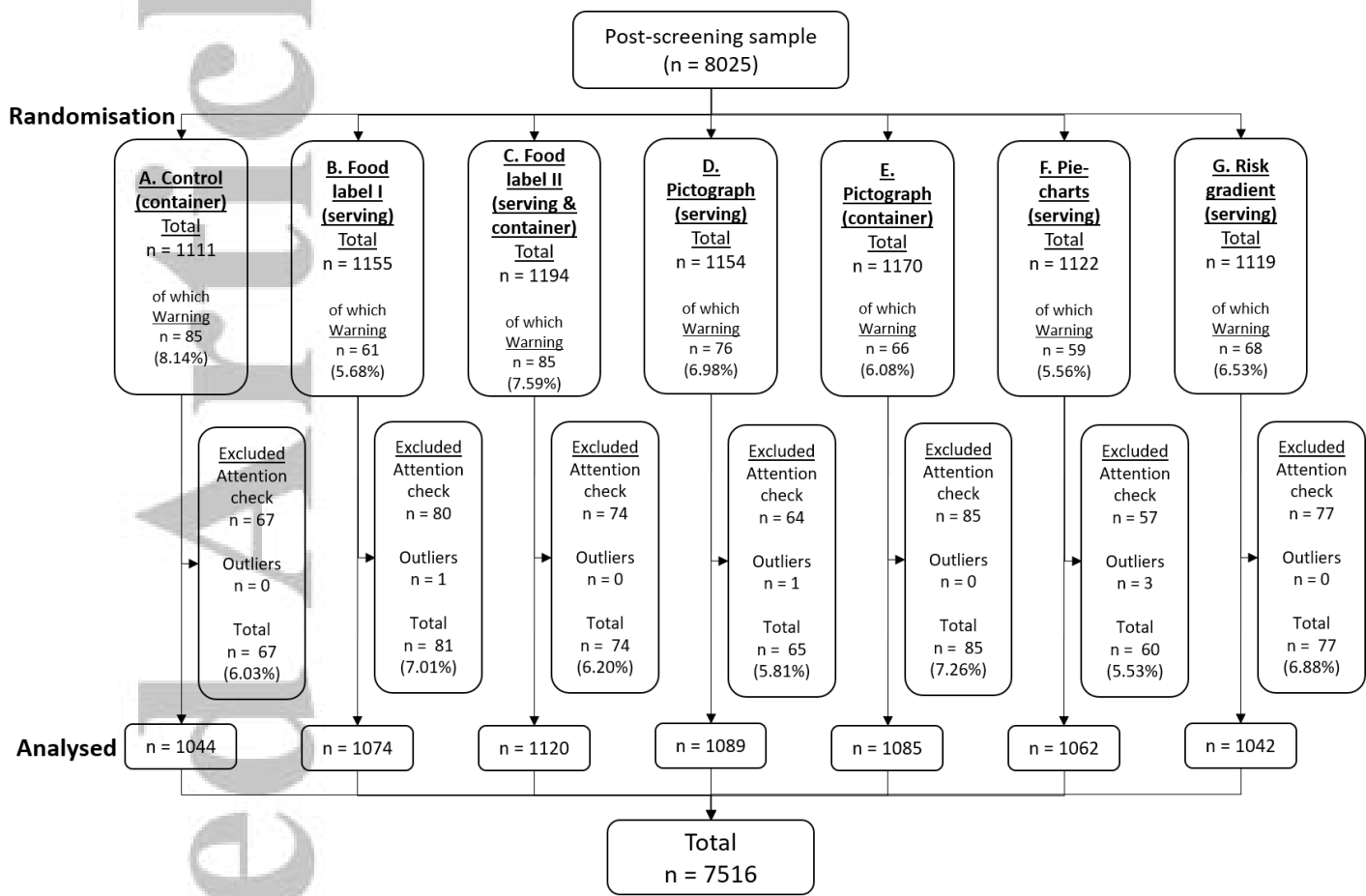
Characteristic	Perceived risk <sup>1</sup>				Motivation to drink <sup>2</sup>				Perception of health-damaging drinking <sup>3</sup>			
	$\beta$ (SE)	95% CIs	p-value		$\beta$ (SE)	95% CIs	p-value		$\beta$ (SE)	95% CIs	p-value	
<b>Treatment (baseline = control)</b>												
Food label serving	-0.04 (0.05)	-0.13 -0.06	0.42		0.14 (0.04)	0.06 0.23	0.001		0.98 (1.44)	-1.85 3.81	0.50	
Food label serving & container	0.02 (0.05)	-0.07 0.12	0.64		0.17 (0.04)	0.08 0.26	<0.001		-0.20 (1.43)	-2.99 2.60	0.89	
Pictograph serving	0.02 (0.05)	-0.07 0.12	0.65		0.17 (0.04)	0.08 0.26	<0.001		-0.04 (1.44)	-2.86 2.77	0.98	
Pictograph container	-0.01 (0.05)	-0.10 0.09	0.89		0.28 (0.04)	0.19 0.36	<0.001		0.91 (1.44)	-1.91 3.72	0.53	
Pie chart serving	0.03 (0.05)	-0.07 0.12	0.57		0.23 (0.04)	0.14 0.32	<0.001		0.93 (1.45)	-2.11 3.57	0.61	
Risk gradient serving	0.05 (0.05)	-0.05 0.15	0.31		0.210 (0.04)	0.12 0.30	<0.001		-1.31 (1.46)	-4.16 1.53	0.37	
<b>Age (baseline = 18-24)</b>												
25-54	-0.02 (0.04)	-0.10 0.06	0.67		0.0 (0.04)	-0.07 0.07	0.99		3.01 (1.18)	0.69 5.33	0.011	
55+	-0.13 (0.04)	-0.21 -0.04	0.003		-0.09 (0.04)	-0.17 -0.02	0.02		5.34 (1.27)	2.85 7.82	<0.001	
<b>Female (baseline = male)</b>												
	0.14 (0.03)	0.09 0.19	<0.001		0.04 (0.2)	-0.00 0.09	0.08		0.83 (0.78)	-0.70 2.36	0.29	
<b>Social grade C2DE (baseline = ABC1)</b>												
	-0.02 (0.03)	-0.08 0.03	0.39		-0.01 (0.03)	-0.06 0.04	0.60		-0.05 (0.83)	-1.68 1.58	0.95	
<b>Ethnicity (baseline = White)</b>												
Black	0.17 (0.09)	0.00 0.34	0.044		0.21 (0.08)	0.06 0.36	0.53		4.33 (2.53)	-0.63 9.28	0.09	
Asian	0.11 (0.07)	-0.03 0.25	0.12		0.24 (0.06)	0.11 0.36	<0.001		-3.14 (2.07)	-7.19 0.91	0.13	
Mixed	-0.10 (0.9)	-0.27 0.08	0.29		0.05 (0.08)	-0.11 0.21	0.007		-0.17 (2.68)	-5.41 5.08	0.95	
Other	-0.5 (0.15)	-0.35 0.24	0.72		0.14 (0.14)	-0.13 0.41	0.32		1.53 (4.50)	-7.30 10.36	0.73	

<b>Region (baseline = North)</b>												
South & East	-0.01 (0.03)	-0.08	0.05	0.65	-0.05 (0.03)	-0.11	0.01	0.10	0.60 (0.96)	-1.27	2.48	0.53
Midlands	0.02 (0.04)	-0.06	0.09	0.63	-0.0 (0.03)	-0.07	0.06	0.93	1.39 (1.11)	-0.79	3.57	0.21
London	-0.01 (0.04)	-0.09	0.08	0.88	0.08 (0.04)	-0.05	0.16	0.04	0.75 (1.28)	-1.76	3.26	0.56
<b>Audit C (numerical, 1- 12)</b>	-0.00 (0.0)	-0.01	0.01	0.63	-0.04 (0.00)	-0.05	-0.03	<0.001	1.89 (0.14)	1.61	2.18	<0.001
<b>Highest education (baseline = none)</b>												
Secondary	0.27 (0.10)	0.08	0.46	0.006	0.03 (0.09)	-0.14	0.21	0.73	3.44 (2.90)	-2.24	9.12	0.24
Post- secondary / Vocational	0.28 (0.10)	0.09	0.46	0.004	-0.01 (0.09)	-0.18	0.16	0.91	2.35 (2.87)	-3.27	7.98	0.41
Undergrad or higher	0.28 (0.10)	0.09	0.47	0.004	0.04 (0.09)	-0.14	0.21	0.67	1.14 (2.90)	-4.55	6.82	0.70
<b>Warning (baseline = no warning)</b>	0.00 (0.05)	-0.10	0.11	0.93	0.07 (0.05)	-0.02	0.17	0.12	-1.28 (1.54)	-4.29	1.74	0.41
<b>Constant</b>	3.60 (0.11)	3.38	3.82	<0.001	3.26 (0.10)	3.05	3.46	<0.001	9.18 (3.38)	2.56	15.80	0.01
<b>R squared</b>		0.01				0.03				0.03		
<b>Sample size</b>		<b>7516</b>				<b>7516</b>				<b>7516</b>		

<sup>1</sup>To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease? From 1 (Not at all likely) to 5 (Extremely likely)

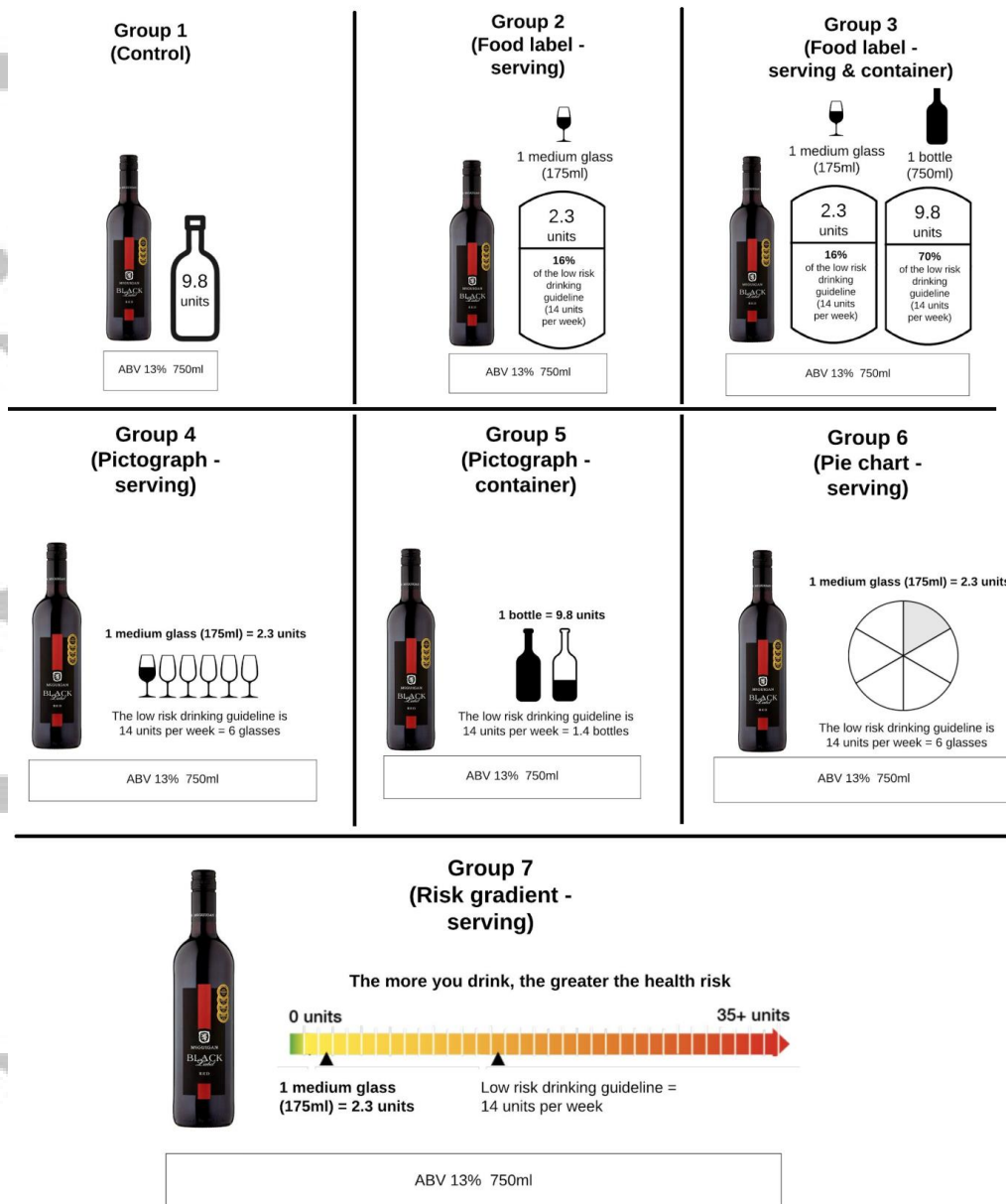
<sup>2</sup> Earlier, you saw the following alcohol label: [beer image #3]. To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less. From 1 (Strongly disagree) to 5 (Strongly agree)

<sup>3</sup> How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health? Free text numeric response

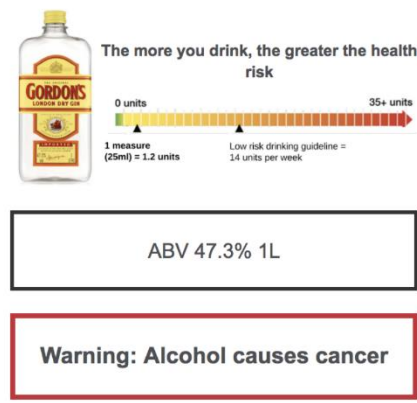
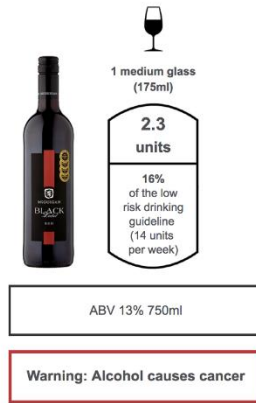
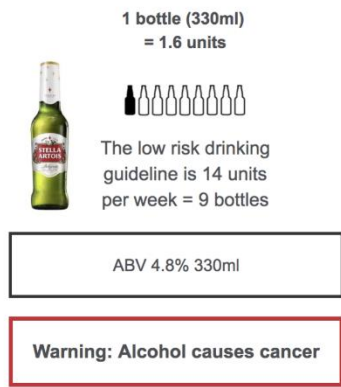


**Figure 1: Trial profile**

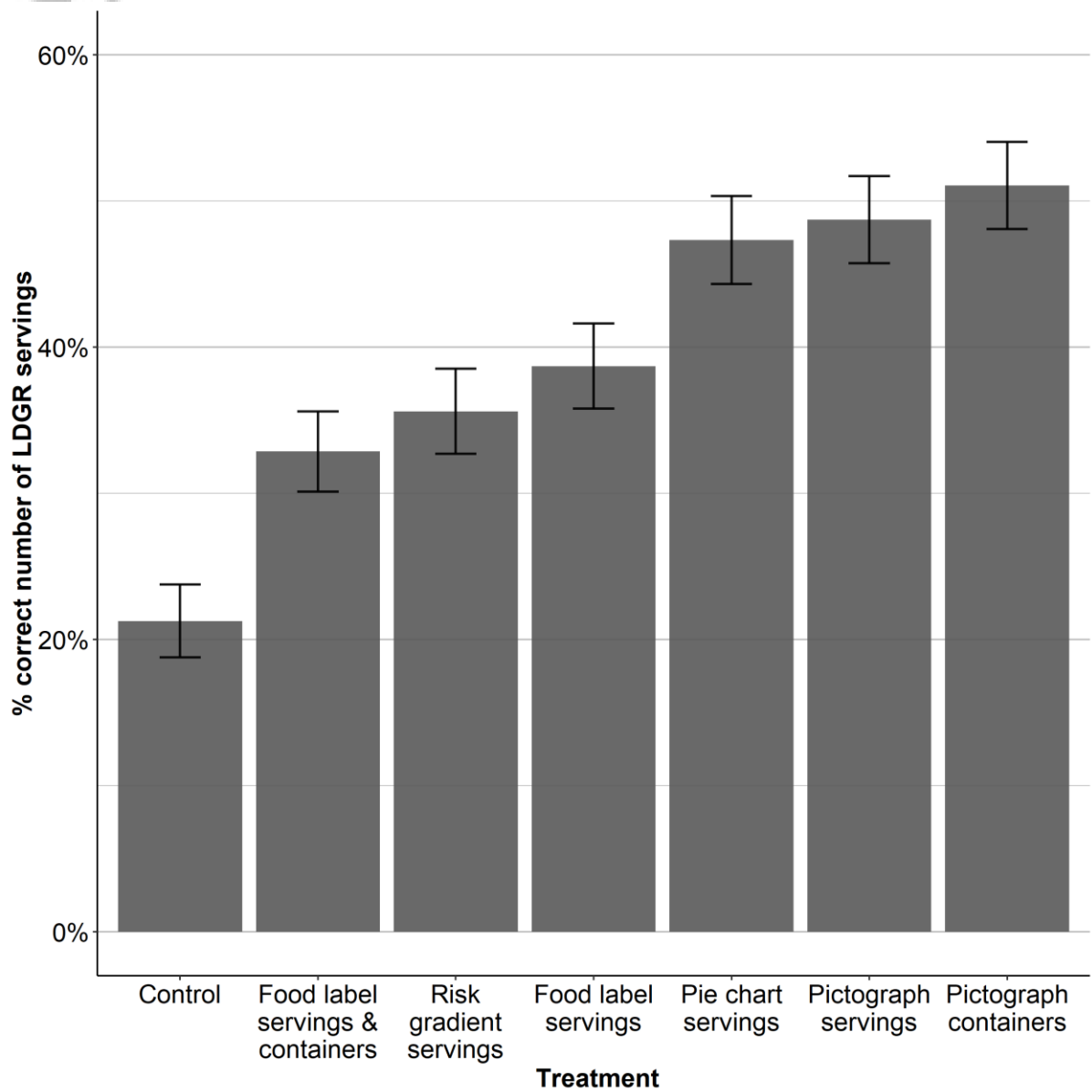




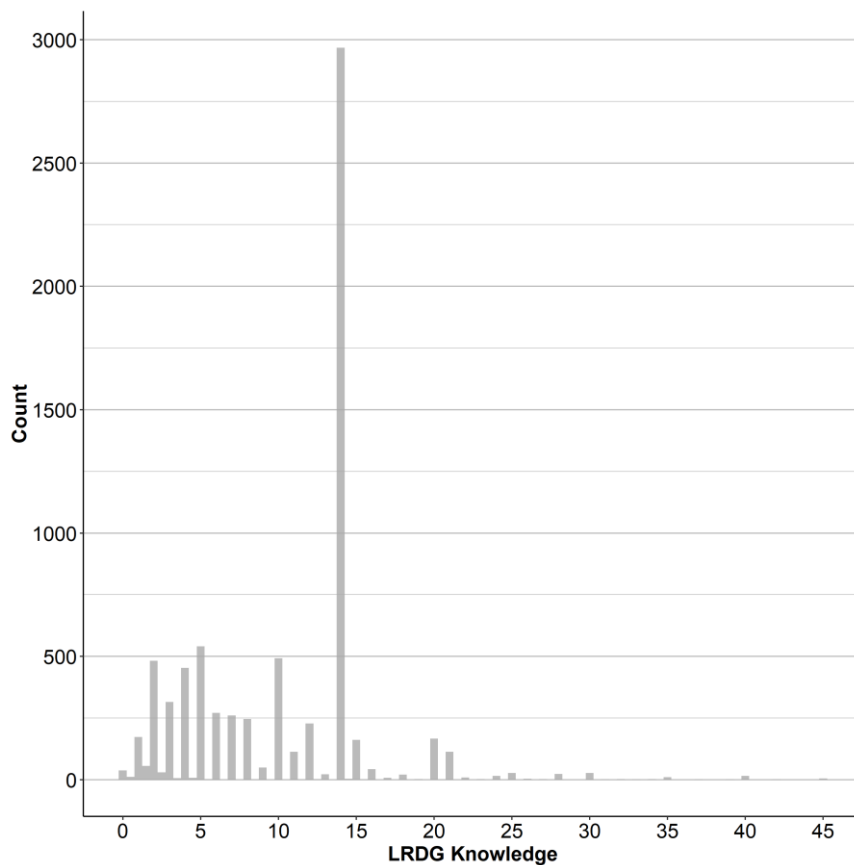
**Figure 2: Example of all seven label designs for one of the wines presented**



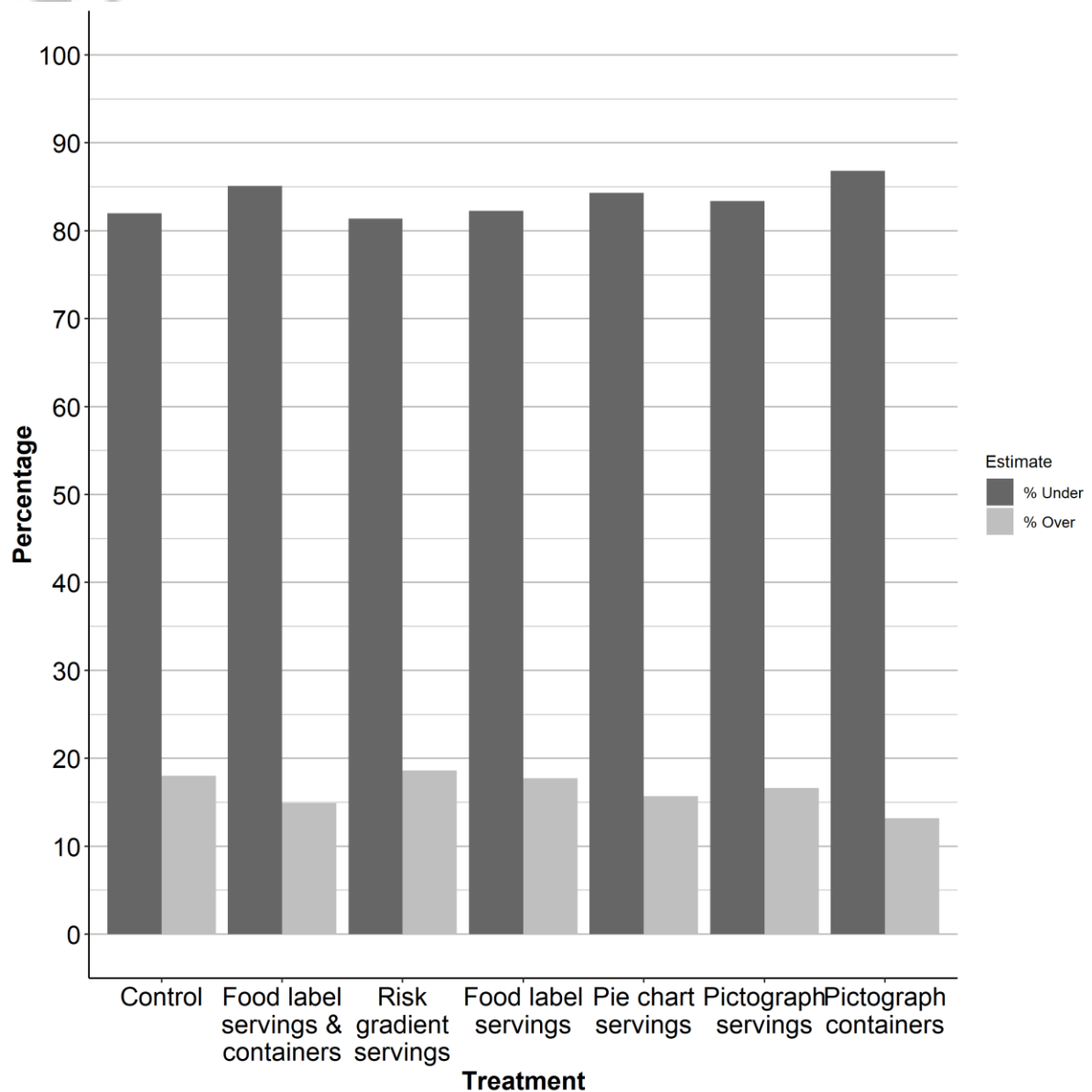
**Figure 3: Example of how the labels with warnings appeared for one beer, one wine, and one spirit label**



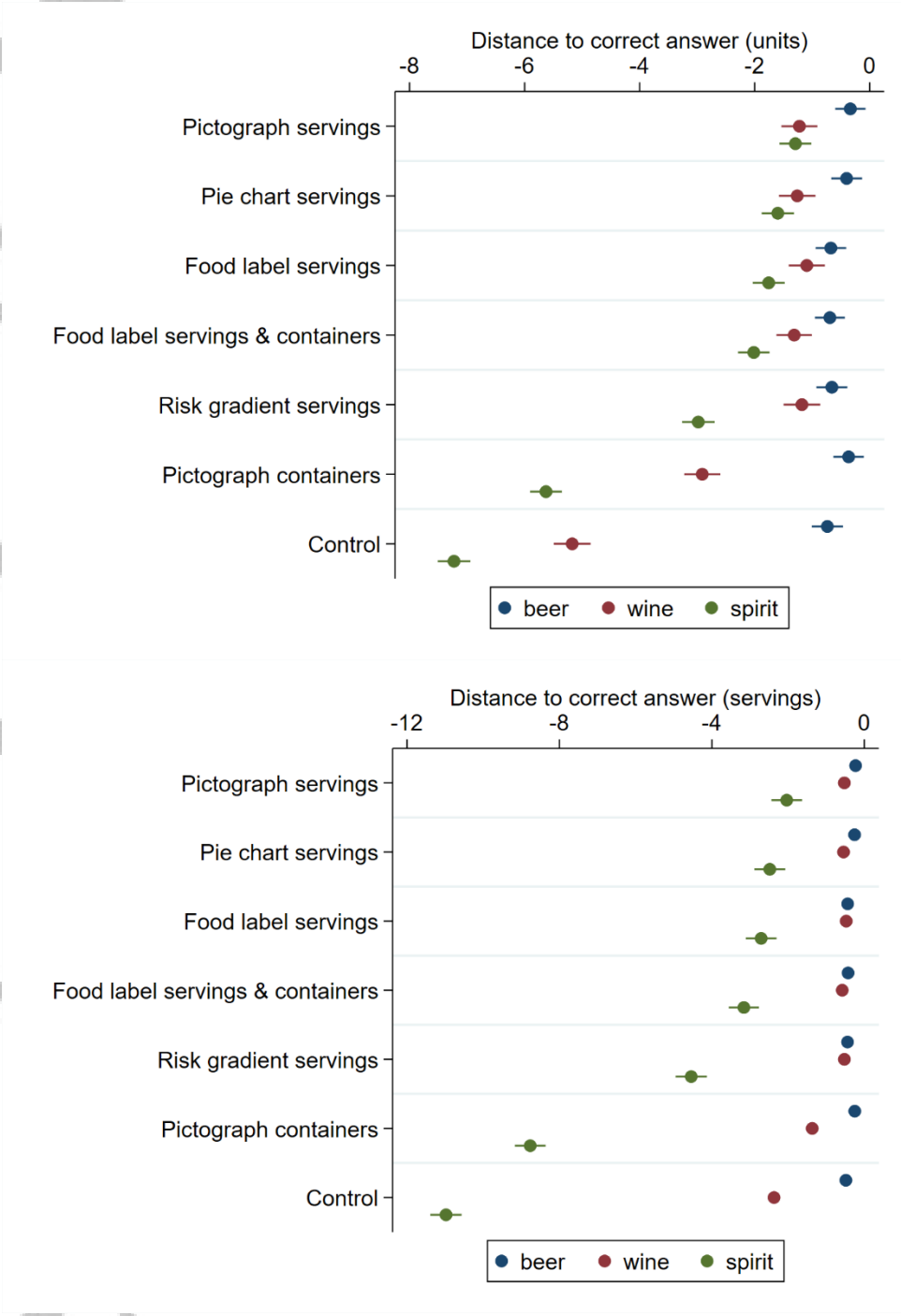
**Figure 4: Bar chart LRDG Knowledge (%) correct with 95% CI bars (by condition)**



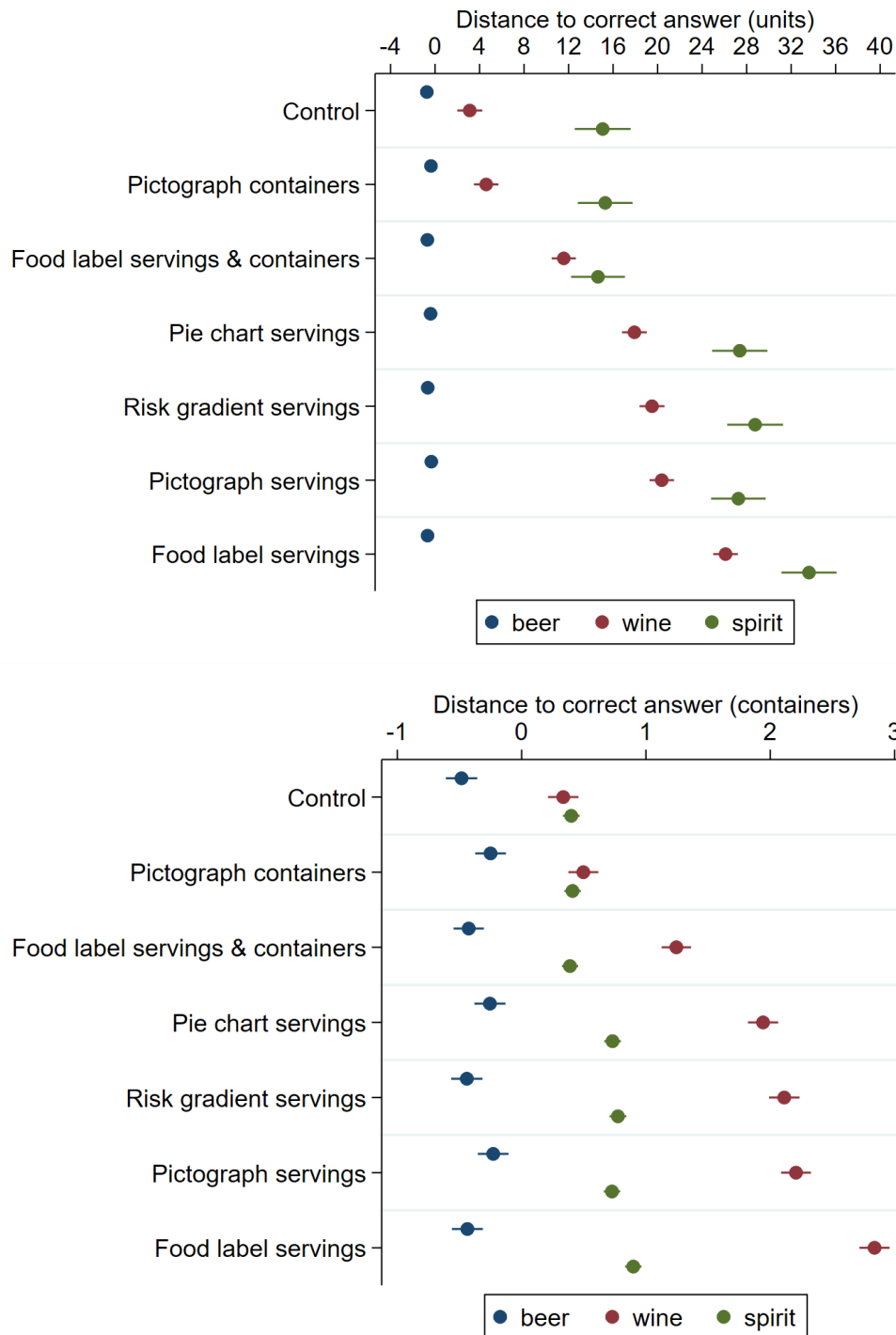
**Figure 5: Distribution of participant responses to LRDG knowledge (LRDG = 14) excluding outlier responses above the 99<sup>th</sup> percentile**



**Figure 6: Participants who gave the incorrect answer to the LRDG, percentage of those who were wrong who under- versus over-estimated.**



**Figure 7: Understanding of the LRDG (servings): How many servings of alcohol can be consumed while remaining under the LRDG? Mean distance from the correct answer in (a) servings and (b) units, ordered from most to least accurate (in terms of aggregate average measure), showing 95% CIs from an OLS regression controlling for demographics**



**Figure 8: Understanding of the LRDG (containers): How many servings of alcohol can be consumed while remaining under the LRDG? Mean distance from the correct answer in (a) containers and (b) units, ordered from most to least accurate (in terms of aggregate average measure), showing 95% CIs from an OLS regression controlling for demographics**

Appendix 2:

Trial protocol and data analysis plan\*

Contents

Protocol.....p.2

Data analysis plan.....p.14

\*This document was written about an experiment and data analysis that was going to be conducted in the future, even though it is written in the past tense,

Accepted



Online Experiment Trial Report:  
Communicating the risk of alcohol consumption: A Predictiv RCT

Trial Number: 2019001

Role	Name
Principal:	Hugo Harper
Policy: Policy QA:	Mark Egan, Kristina Londakova, Max Kroner Dale, Helen Brown Toby Park
Research: Research QA:	Mark Egan, Abigail Mottershaw Martin Sweeney
<b>Partner Organisation: Public Health England</b>	
Partner Lead	Maria Smolar <a href="mailto:maria.smolar@phe.gov.uk">maria.smolar@phe.gov.uk</a>

Accepte

### Partner information

Public Health England (PHE) is an executive agency of the Department of Health and Social Care. Its purpose is to make the public healthier by providing evidence-based scientific expertise and support to organisations and the broader public.

In October 2018, PHE's Alcohol Team commissioned BIT to design and test "*visual or language risk-based messages for the general drinking public that helps people understand the risks that alcohol poses to their health*".

To inform the design of these messages, BIT conducted a [rapid evidence review](#) (Nov 2018 - Jan 2019) to summarise findings from existing research, ran a 10-person focus group (Dec 2018) on the topic of alcohol labels, and consulted with key stakeholders at PHE and academic experts at the Tobacco and Alcohol Research Group at Bristol University.

This exploratory work is now followed by a large Predictiv randomised controlled trial (RCT) which tests new risk messaging, co-designed by BIT and PHE.

Accepted

## 1.1 Background

In 2016, the United Kingdom Chief Medical Officers (CMOs) published low risk drinking guidelines, aimed at the general public, which read "To keep health risks from alcohol to a low level it is safest not to drink more than 14 units a week on a regular basis."<sup>1</sup>

The following year, the Department of Health (DH) recommended that the CMOs' guidelines be communicated to the public using the following visual prompts.<sup>2</sup> While DH recommended that these three guidelines be grouped together clearly and legibly on the primary packaging of alcohol products, it is not compulsory for drinks producers to do this.

**Figure 1: Department of Health guidelines for communicating alcohol risk.**



These recommendations updated the previous guidelines, published in 1995, which recommended a maximum intake of 21 units of alcohol per week for men and 14 for women.<sup>3</sup> The updated 2016 guidelines revised these to 14 units per week for both men and women in light of new high quality evidence linking alcohol with worse health outcomes than previously realised, particularly the causal relationship between alcohol and cancer.<sup>4</sup>

Despite the health risks involved with alcohol consumption, there are three key aspects to consider when aiming to understand the effectiveness of the new guidelines. These are:

1. **Lack of awareness.** Recent representative surveys of the adult British population have found that while most people have heard of alcohol units

1

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/545937/UK\\_CMOs\\_report.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/545937/UK_CMOs_report.pdf)

2

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/602132/Communicating\\_2016\\_CMO\\_guidelines\\_Mar\\_17.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/602132/Communicating_2016_CMO_guidelines_Mar_17.pdf)

3

[https://webarchive.nationalarchives.gov.uk/20130105043158/http://www.dh.gov.uk/prod\\_consum\\_dh/groups/dh\\_digitalassets/@dh/@en/documents/digitalasset/dh\\_4084702.pdf](https://webarchive.nationalarchives.gov.uk/20130105043158/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4084702.pdf)

4

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/545911/GovResponse2.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/545911/GovResponse2.pdf)

and of the existence of the alcohol guidelines, only 8-19% actually knew that the low risk drinking guideline was 14 units per week.<sup>5,6</sup>

2. **Lack of understanding.** The guidelines are based around alcohol 'units', and a broad review of alcohol units (or equivalent metrics) in multiple countries concluded that communicating the concept is inherently difficult given the wide variation in the strengths of different alcohols and the different amounts of alcohol different people tend to pour.<sup>7</sup>
3. **Lack of Interest.** A study involving 12 focus groups in England and Scotland found that many respondents thought the official drinking guidelines were unrealistically low or irrelevant to their own lives.<sup>8</sup> That said, a 2018 survey of 450 drinkers in the UK found that 91% supported providing alcohol unit information on drinks, suggesting that there is public support for the general idea of risk guidelines.<sup>9</sup>

## 1.2 Project Aims

This research aims to test whether enhanced alcohol labels can improve awareness and understanding of alcohol-related health risk.

## 1.3 Predictiv

Predictiv is an online platform for running behavioural experiments built by the Behavioural Insights Team. It enables governments and other organisations to run randomised controlled trials (RCTs) with an online population of participants, and to experiment whether new policies and interventions work before they are deployed in the real world. Predictiv provides access to a large international panel, including more than 200,000 individuals in the UK and 1,000,000 in the US, as well as the functionality to run a range of online experiments. More information on the methodology behind Predictiv, including payments, randomisation, recruitment, data storage and ethics can be found [here](#). This trial follows these standard procedures.

<sup>5</sup> Rosenberg (2017) New national alcohol guidelines in the UK: public awareness, understanding and behavioural intentions

<sup>6</sup> [http://12coez15v41j2cf7acjzaodh.wpengine.netdna-cdn.com/wp-content/uploads/2018/11/AHA\\_How-we-drink-what-we-think\\_2018\\_FINAL.pdf](http://12coez15v41j2cf7acjzaodh.wpengine.netdna-cdn.com/wp-content/uploads/2018/11/AHA_How-we-drink-what-we-think_2018_FINAL.pdf)

<sup>7</sup> Kerr & Stockwell (2012) Understanding standard drinks and drinking guidelines

<sup>8</sup> Lovatt M, Eadie D, Meier PS, Li J, Bauld L, Hastings G, Holmes J. 2015. Lay epidemiology and the interpretation of low-risk drinking guidelines by adults in the United Kingdom. *Addiction*.

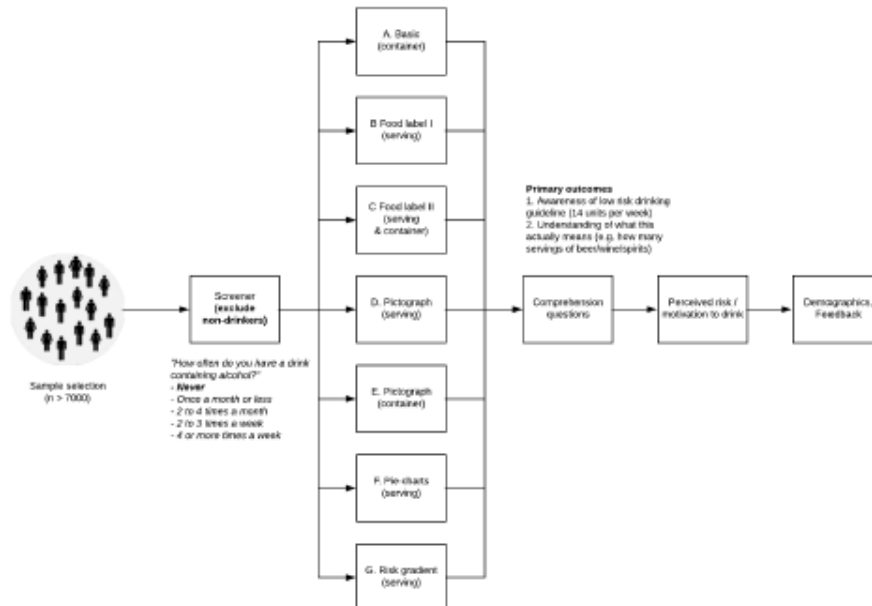
<sup>9</sup> [https://s3.eu-west-](https://s3.eu-west-2.amazonaws.com/files.alcoholchange.org.uk/documents/FinalReport_0150.pdf?mtime=20181110145642)

[2.amazonaws.com/files.alcoholchange.org.uk/documents/FinalReport\\_0150.pdf?mtime=20181110145642](https://s3.eu-west-2.amazonaws.com/files.alcoholchange.org.uk/documents/FinalReport_0150.pdf?mtime=20181110145642)

## 1.4 Experimental Design and Procedures

This trial was a 7-arm RCT with 1000 people per arm, as shown in Figure 2.

Figure 2: Experiment design



*This is a simplified presentation of the experiment design. Within each of the 7 arms (A-G), we tested versions of each label with and without an additional warning. This meant there were a total of 14 stimuli (7 label types X 2 warning types) participants could see. However, we did not conduct primary analysis across all 14 cells.*

This experiment was conducted on Predictiv. Participants could choose to participate in this experiment through the panel survey website on which they were registered. They were then taken through several stages:

- **Material stage:** Participants were randomly allocated into one of 7 main intervention arms, each of which showed a different type of alcohol label.
- **Comprehension questions:** Participants were asked questions about their awareness and understanding of alcohol risk, and their perceived risk of alcohol consumption and motivation to drink.

- **Demographic questions:** Participants answered demographic questions and provided feedback.

#### Pre-test

Before the full experiment launched, a pilot involving the same test material was conducted on 200 people. This was to ensure that the experiment was working as intended. Specifically, we tested:

- 1) How long did the experiment take to complete on average?

*If the median participant took longer than 10 minutes to complete the experiment, we would reduce the number of questions / amount of stimuli in the experiment.*

- 2) Are there interaction effects between the treatment arms and the labels?

*This trial tested the effect of warning messages on people's motivation to drink and their perception of alcohol risk. Due to resource constraints, it did this by testing warning messages within treatment arms rather than testing the warnings in their own separate treatment arms (e.g. within each treatment arm involving 1000 people, we randomly assigned 500 to see a version of the alcohol label with an additional warning and assigned 500 to see a version without the warning),.*

*This approach incurred a risk of interaction effects, e.g. it was possible that showing participants an alcohol label bundled with a warning would cause them to provide meaningfully different responses to the primary outcomes (i.e. those measuring awareness and understanding of the low risk drinking guideline) compared to those who just saw an alcohol label without a warning.*

*We therefore used the pilot data (n=200) to examine whether the group of approx. 100 people who saw alcohol labels with a warning provided significantly different responses (at  $p < 0.1$ ) on the primary outcomes compared to the group of approx. 100 people who saw the labels without a warning.*

*If there was a large difference between these groups (note the difference would have to be large to be detectable at  $p < 0.1$  in a sample of 200 people), then we would not use the within-arm approach to test the efficacy of warnings in the main experiment.*

### 1.5 Treatments and randomisation

After viewing an introduction page participants were randomly assigned to see one of 7 types of alcohol label.<sup>10</sup> For each type of label, participants first saw a page containing three example beers, then a page with three example wines, then a page with three example spirits. Figure 3 shows examples of how these materials appeared to people in three different groups.

---

<sup>10</sup> Participants were randomly assigned to a treatment at an individual level. When a participant entered the experiment, they were given a random number representing an intervention. Depending on the number assigned, they were taken through a separate path in the experiment that corresponded with a specific intervention (e.g. the 'risk axis' labels). The random number was stored in the data output and used for data analysis to assess the interventions' impact on the outcome variables.

Figure 3: Example alcohol labels.

Control group [beer]



Food label (serving) group [wine]



Pictograph (container) group [spirits]








Acc



Table 1 briefly describes all 7 types of alcohol labels tested in the experiment, all of which included the product's alcohol-by-volume (ABV %) and volume (e.g. '750ml') at a minimum. These labels, and the user-journey for a person in the control group, are [shown in full here](#).

**Table 1: Overview of different conditions in the test.**

#	Condition	Description	Example	N
1	Control (container)	<p>Number of units per bottle.</p> <p><i>Note this is the 'Responsibility Deal' design currently recommended for use by the alcohol industry in the UK.</i></p>	 <p>ABV 4.8% 250ml</p>	1000
2	Food label (serving)	<p>Number of units per serving, expressed as a percentage of the low-risk amount.</p>	 <p>ABV 4.8% 330ml</p>	1000

3	Food label (serving & container)	<p>Number of units per serving <i>and</i> the whole container, both expressed as a percentage of the low-risk amount.</p> <p><i>Note. For beer, this image only describes the number of units in 1 bottle because for beer, a serving and a container is the same thing.</i></p>		1000
4	Pie-chart (serving)	<p>The number of units per serving, as a share of a pie-chart indicating the low risk amount.</p>		1000
5	Pictograph (serving)	<p>The number of units per serving, using pictographs indicating the low risk amount.</p>		1000

6	Pictograph (container)	The number of units per container, using pictographs indicating the low risk amount.  <i>Note. For beer, treatments 5 and 6 present the same image (because 1 bottle is both a serving and a container for beer)</i>		1000
7	Risk gradient (serving)	The number of units per bottle as on a visual axis representing increasing alcohol-related risk.		1000
Total				7000

Within each arm, we also randomly assigned people to see one of two 'warning' conditions (e.g. for 1000 people in each arm, two groups of 500 were assigned to see one of the below):

- (i) **No warning.** This baseline group saw only the labels shown in Table 1, with no additional warning.
- (ii) **CMO warning.** This group saw the labels shown in Table 1 *and* the below warning.

**Warning: Alcohol increases your risk of cancer**

Figure 4: Example alcohol labels with additional warning.



This meant there were 14 possible stimuli participants could be randomly assigned to see (7 labels X 2 warnings). However, as described below, we did not conduct primary analysis across all 14 cells.

## 1.6 Participant pool and eligibility

We recruited 7000 participants (1000 per arm) after attrition.

Eligible participants were:

- drinkers of alcohol, and
- adults living in England.

Alcohol consumption eligibility was determined using the following screener question:

"How often do you have a drink containing alcohol?"

- *Never*
- *Once a month or less*
- *2 to 4 times per month*
- *2 to 3 times per week*
- *4 or more times per week*

Participants who answered 'never' did not proceed to take part in the study.

Age and geographic eligibility (e.g. people aged 18 and over living in England) was identified using the information already on file for these variables for all Predictiv participants. In

addition, we aimed to select a sample which was (approximately) nationally representative of the adult population of England in terms of age, gender and region. The sample characteristics we sought to achieve were:<sup>11</sup>

- Gender: 49% Male, 51% Female
- Age: 12% age 18 - 24, 51% 25 - 54, 37% 55 and over
- Location: 28% North England, 37% South and East England, 19% Midlands, 16% London.

---

<sup>11</sup> Sources for England figures:

Gender: <https://www.nomisweb.co.uk/census/2011/KS101EW/view/2092957699?cols=measures>

Age: <https://www.nomisweb.co.uk/census/2011/KS102EW/view/2092957699?cols=measures>

Region: <https://www.nomisweb.co.uk/census/2011/ks101ew>

## 1.7 Outcome measures

Table 2 lists our 6 outcome measures (3 primary outcomes and 3 secondary ones).

**Table 2: Outcome measures.**

# (outcome type)	Measure	Question text	Coding
1 (Primary)	<b>Awareness of low risk drinking guideline</b> (14 units per week)	<p><i>"The government's low risk drinking guideline recommends that people not regularly drink more than a certain number of alcohol units per week.</i></p> <p><i>What do you think the low risk drinking guideline is?"</i></p> <p>Participants answer by entering numeric free text.</p>	Binary variable where: 1 = participant says '14' units 0 = participant does not say this
2 (Primary)	<b>Understanding of low risk drinking guideline in terms of servings</b> (i.e. whether participants can accurately estimate how many bottles of beer / glasses of wine / shots of spirits it takes to reach this guideline).	<p><i>"How many [bottles of this beer (330ml) / medium-size (175ml) glasses of this wine / single shots (25ml) of this drink] do you think it takes to get to 14 units?"</i></p> <p>Participants answer by entering numeric free text; they are asked this type of question 2 times for beer, 2 times for wine, and 2 times for spirits (6 questions total).</p>	<p>For each of the 6 items, we measured <i>distance to the correct response</i> (e.g. if the correct answer was "6 bottles of beer", a participant who said '6' got a score of 0; a participant who said '5' got a score of -1; a participant who said '10' got a score of 4).</p> <p>We then took the average of these 6 items to create a single composite measure.</p>
3 (Primary)	<b>Understanding of low risk drinking</b>	<i>"How many [bottles of this beer (330ml) / bottles (750ml)</i>	For each of the 6 items, we measured <i>distance</i>

	<b>guideline in terms of containers</b> (i.e. whether participants can accurately estimate how many bottles of beer / bottles of wine / bottles of spirits it takes to reach this guideline).	<i>glasses of this wine / bottles (700ml) of this drink] do you think it takes to get to 14 units?"</i>  Participants answer by entering numeric free text; they are asked this type of question 2 times for beer, 2 times for wine, and 2 times for spirits (6 questions total).	to the correct response (e.g. if the correct answer was "6 bottles of beer", a participant who said '6' got a score of 0; a participant who said '5' got a score of -1; a participant who said '10' got a score of 4).  We then took the average of these 6 items to create a single composite measure.
4 (Secondary)	<b>Perceived risk of own alcohol consumption.</b>	<i>"To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease?"</i>	Categorical variable coded as: 1 = Not at all likely, 2 = Not very likely, 3 = Somewhat likely, 4 = Quite likely, 5 = Extremely likely
5 (Secondary)	<b>Motivation to drink.</b>	<i>"Earlier, you saw the following alcohol label: [beer image #3]. To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less."</i>	Categorical variable coded as: 1 = Strongly disagree 2 = Disagree 3 = Neither agree nor disagree 4 = Agree 5 = Strongly agree
6 (Secondary)	<b>Subjective perception of 'high risk' drinking.</b>	<i>"Lastly, we are interested in your opinion about the following: How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health?"</i>	Continuous variable

		Participants answer by entering numeric free text.	
--	--	--	--

The first primary outcome measured **awareness of the low risk drinking guideline**. This simply examined what proportion of people knew that low risk guideline was 14 units per week.

The other primary outcome measured **understanding of this low risk drinking guideline**. In the real world, simply knowing that the guideline is '14 units per week' might not be helpful to people if they are not able to take the next step of contextualising this in terms of their own drinking (similar to how telling people they should eat 2000 calories per day might not be helpful if people do not also understand how many calories there are in different food products).

To measure this outcome, we showed people a series of alcohol products, then asked them how many servings/containers they thought they could have of the product (e.g. how many bottles of beer, how many glasses of wine, how many shots of spirits) before reaching the 14 unit guideline. This outcome was designed to be similar to the experience a person might have of standing in an aisle of a supermarket, looking at different alcohol products, as they try to quickly assess how much of a certain product they could have before hitting the guideline.

We measured this outcome using the following 10 questions (see [slides 15-23](#) for an example of how these appeared to participants in the experiment):

1. **Beer I.** How many bottles of this beer (330ml) could you have before reaching 14 units? *[Participants answered this while viewing beer image #2 they had previously been randomly assigned to see].*
2. **Beer II.** How many cans of this beer (568ml) could you have before reaching 14 units? *[Participants answered this while viewing beer image #3 they had previously been randomly assigned to see].*



3. **Wine I (serving)**. How many medium-size (175ml) glasses of this wine could you have before reaching 14 units? *[Participants answered this while viewing wine image #1 they had previously been randomly assigned to see].*
4. **Wine II (serving)**. How many medium-size (175ml) glasses of this wine could you have before reaching 14 units? *[Participants answered this while viewing wine image #3 they had previously been randomly assigned to see].*
5. **Spirits I (serving)**. How many single shots (25ml) of this drink could you have before reaching 14 units? *[Participants answered this while viewing spirits image #1 they had previously been randomly assigned to see].*
6. **Spirits II (serving)**. How many single shots (25ml) of this drink could you have before reaching 14 units? *[Participants answered this while viewing spirits image #2 they had previously been randomly assigned to see].*
7. **Wine III (container)**. How many bottles (750ml) of this wine could you have before reaching 14 units? *[Participants answered this while viewing wine image #2 they had previously been randomly assigned to see].*
8. **Wine IV (container)**. How many bottles (750ml) of this wine could you have before reaching 14 units? *[Participants answered this while viewing wine image #1 they had previously been randomly assigned to see].*
9. **Spirits III (container)**. How much of a bottle or whole bottles (700ml) of this drink could you have before reaching 14 units? *[Participants answered this while viewing spirits image #2 they had previously been randomly assigned to see].*
10. **Spirits IV (container)**. How much of a bottle or whole bottles (1L) of this drink could you have before reaching 14 units? *[Participants answered this while viewing wine image #3 they had previously been randomly assigned to see].*

Accuracy for these 10 questions was measured using *distance from the correct response*. In other words, for each question, there was a technically correct answer which could be calculated using the product's ABV and volume information (those correct answers are [listed here](#)). We did not expect many people in the experiment to get the technically correct answer to these questions. Rather, we expected that many people would get *approximately* the right answer (e.g. estimating they could have 8 bottles of beer when the technically correct answer was 8.4 bottles).

Before conducting analysis on this question, we first collapsed people's responses into two composite variables:

1. **Accuracy (servings).** This took the average of items 1-6 described above (i.e. their average accuracy estimates for how many bottles of beer, glasses of wine, and shots of spirits they could have before reaching 14 units).
2. **Accuracy (containers).** This took the average of items 1, 2, 7, 8, 9, and 10 described above (i.e. their average accuracy estimates for how many bottles of beer, bottles of wine, and bottles of spirits they could have before reaching 14 units). Note this means we included people's estimates for the 'beer' questions for both the 'accuracy (servings)' and 'accuracy (containers)' outcomes; this is because we ask people to estimate how many bottles of beer they think they can have, and for beer a bottle is both a serving and a whole container (whereas this is not the case for wine or spirits).

We separated these responses into two separate composite variables, rather than a single composite variable averaging people's responses across all ten items, for two main reasons:

(i) **Realism.** From the focus group we ran as part of this project, it became apparent that different people prefer to contextualise and understand their own drinking in different ways - some think in terms of servings (e.g. how many glasses of wine or shots of spirits they can have before reaching the 14 unit guideline), others find it easier to think in terms of bottles (e.g. how many bottles of wine or spirits they can have). We therefore considered that the role of an alcohol label should be to help people contextualise their drinking in terms of servings *and/or* containers, but not necessarily both. In the real world, people do not necessarily need to understand how to contextualise their drinking in terms of both servings (e.g. knowing they can have 5 glasses of wine) *and* containers (e.g. knowing they can have 1.3 bottles of wine) - understanding either one should allow them, in most circumstances, to figure out how much they can drink to stay within the low risk drinking guideline.

(ii) **Fairness.** Some of the alcohol labels tested in this experiment focused on helping people understand how many servings (e.g. glasses of wine) they could have before reaching the 14 unit limit; others focused on communicating on how many containers they could have (e.g. bottles of wine); another tried to communicate both pieces of information (e.g. how many glasses *and* bottles of wine they could have).

Our view was that a fair test of the labels should separately examine how effective they were at helping people understand how many servings and containers they could have before reaching the 14 unit guideline.

An example may help explain our rationale. Consider the image to the right - this is an alcohol label shown to people in the 'Pictograph (serving)' group. We might

expect people who saw this image to do reasonably well when asked 'How many medium-size (175ml) glasses of this wine could you have before reaching 14 units?', since this can be worked out quickly and directly by looking at the image (the answer is 7 glasses). If people answered this question accurately, this would mean that the label had done its job at providing people with a practical visual aid for contextualising their own drinking - in this case that aid helps people realise how many *servings* they can have, rather than containers.



During the experiment, people who saw this same image were also asked "How many bottles (750ml) of this wine could you have before reaching 14 units?". It is possible to work out the correct answer to this question using the information in the image, but it is a very complicated calculation to make, particularly since participants are instructed not to use a calculator. We therefore expected that people who saw that image might do relatively poorly on this question.

The opposite situation was true for other treatment materials. For example, the image to the right makes it relatively simple to answer the question "How much of a bottle or whole bottles (700ml) of this drink could you have before reaching 14 units?" - again this information is directly provided. However, this image does not make it easy to answer "How many single shots (25ml) of this drink could you have before reaching 14 units?" In other words, the label helps people contextualise their drinking in terms of containers, but not in terms of servings.



Of the 7 labels tested in this experiment, only the "food label (serving+container)" group attempted to give people a quick guide for understanding of how many servings *and* containers they could have of a certain product before reaching the guideline. An example of this label is shown in the image to the right.



Table 3 shows the control variables used in our analysis.

**Table 3: Control measures.**

Measure	Definition	Coding
Treatment	Treatment assignment	Categorical variable: 0 = Control 1 = Treatment 1 2 = Treatment 2 3 = Treatment 3 4 = Treatment 4 5 = Treatment 5 6 = Treatment 6
Gender	“What is your gender?” *	Categorical variable: 0 = Male 1 = Female
Age	“What is your age?” *	Categorical variable: 0 = 18-24 1 = 25-54 2 = 55+
Household income	“What is your current annual household income before taxes?” *	Categorical variable based on median income in UK: 0 = < £30,000 1 = >= £30,000
Location	“In which region do you live?” * ; Original variable has 12 levels. (NUTS1).	Categorical variable: 0 = London 1 = North East, North West, Yorkshire & Humber 2 = East of England, South East, South, West 3 = East Midlands, West Midlands
Education level	“What is the highest education level that you have achieved?”	Categorical variable: 0 = None 1 = Secondary school 2 = Post-secondary / vocational 3 = Undergraduate or above

<p>Alcohol use disorders identification test consumption score (<a href="#">AUDIT C</a>)</p>	<p>Three-item scale to measure alcohol consumption:</p> <ol style="list-style-type: none"> <li>1. How often do you have a drink containing alcohol?</li> <li>2. How many units of alcohol do you drink on a typical day when you are drinking?</li> <li>3. How often have you had 6 or more units if female, or 8 or more if male, on a single occasion in the last year?</li> </ol> <p><i>Note 1: Item 1 is asked at the start of the experiment and is used as a screener question. Items 2 and 3 are asked towards the end.</i></p> <p><i>Note 2: Items 2 &amp; 3 are answered with reference to a visual guide.</i></p>	<p>Continuous variable scored 0-12 (where higher scores mean greater alcohol consumption).</p>
<p>Social grade</p>	<p>“Could you tell us what the profession of the chief income earner in your household is?”</p>	<p>0 = upper middle class ('A' = Higher managerial/ professional/ administrative)  1 = middle class ('B' = Intermediate managerial/ professional/ administrative)  2 = lower middle class ('C1' = Supervisory or clerical/ junior managerial/ professional/ administrative)  3 = skilled working class ('C2' = Skilled manual worker)  4 = working class ('D' = Semi or unskilled manual work)  5 = non working ('E' Casual worker – not in permanent employment, Housewife/ Homemaker, Retired and living on state pension, Unemployed or not working due to long-term sickness, Full-time career of other household member, Student)</p>
<p>Smoking</p>	<p>“Do you smoke cigarettes at all nowadays?”</p>	<p>0 = no  1 = yes</p>

Ethnicity	“What is your ethnic group? Choose one option that best describes your ethnic group or background.”	0 = White 1 = Mixed/multiple ethnic groups 2 = Asian / Asian British 3 = Black/African/Caribbean/Black British 4 = Other ethnic group
* Participants are automatically profiled on standard demographic characteristics (age, gender, location, income), which means that this information does not need to be solicited in the experiment.		

## 1.8 Analysis strategy

### Primary outcomes

1. **Awareness of low risk drinking guideline (LRDG).** Participants were asked "The government's low risk drinking guideline recommends that people not regularly drink more than a certain number of alcohol units per week. What do you think the low risk drinking guideline is?"

As described in Table 2, this outcome was coded as a binary measure where 1 = participant said "14" units per week", 0 = any other number of units.

2. **Understanding of LRDG (serving).** Participants answered 6 questions where they were shown an image of an alcohol label (of the same type they had already seen earlier in the experiment), and then asked: "How many [bottles of this beer (330ml) / medium-size (175ml) glasses of this wine / single shots (25ml) of this drink] do you think it takes to get to 14 units?". Their responses were averaged into a single composite variable in the manner described in Table 2.
3. **Understanding of LRDG (container).** Participants answered 6 questions where they were shown an image of an alcohol label (of the same type they had already seen earlier in the experiment), and then asked: "How many [bottles of this beer (330ml) / bottles (750ml) glasses of this wine / bottles (700ml) of this drink] do you think it takes to get to 14 units?". Their responses were averaged into a single composite variable in the manner described in Table 2.

The 'Awareness' outcome was examined using the following Logit model:

$$Y_i^{\text{Awareness of LRDG}} = \alpha + \varphi_{\text{Condition [label]}} + \varphi_{\text{Gender}} + \beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + e_i$$

Where:

$Y_i^{\text{Awareness of LRDG}}$  was the outcome measure for each individual  $i$ .

$\alpha$  was the constant.

$\text{Condition [label]}_i$  was a vector of binary variables indicating which of the 7 label types the participant was randomly assigned to see.

## 1.8 Analysis strategy

### Primary outcomes

1. **Awareness of low risk drinking guideline (LRDG).** Participants were asked "The government's low risk drinking guideline recommends that people not regularly drink more than a certain number of alcohol units per week. What do you think the low risk drinking guideline is?"

As described in Table 2, this outcome was coded as a binary measure where 1 = participant said "14" units per week", 0 = any other number of units.

2. **Understanding of LRDG (serving).** Participants answered 6 questions where they were shown an image of an alcohol label (of the same type they had already seen earlier in the experiment), and then asked: "How many [bottles of this beer (330ml) / medium-size (175ml) glasses of this wine / single shots (25ml) of this drink] do you think it takes to get to 14 units?". Their responses were averaged into a single composite variable in the manner described in Table 2.
3. **Understanding of LRDG (container).** Participants answered 6 questions where they were shown an image of an alcohol label (of the same type they had already seen earlier in the experiment), and then asked: "How many [bottles of this beer (330ml) / bottles (750ml) glasses of this wine / bottles (700ml) of this drink] do you think it takes to get to 14 units?". Their responses were averaged into a single composite variable in the manner described in Table 2.

The 'Awareness' outcome was examined using the following Logit model:

$$Y_i^{\text{Awareness of LRDG}} = \alpha + \varphi_{\text{Condition}} [\text{label}]_i + \varphi_{\text{Gender}} + \beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + e_i$$

Where:

$Y_i^{\text{Awareness of LRDG}}$  was the outcome measure for each individual  $i$ .

$\alpha$  was the constant.

$\text{Condition} [\text{label}]_i$  was a vector of binary variables indicating which of the 7 label types the participant was randomly assigned to see.



$\varphi_{\text{individual}}$  was a vector of binary variables indicating each participants' (1) age category, (2) social grade, (3) ethnicity, (4) education, and (5) region, coded as described in Table 3.

*Gender* and *Audit C score* were binary and continuous variables respectively and were coded in the manner described in Table 3.

$e_i$  was the heteroscedasticity-robust error term.

The 'Understanding' outcomes were examined using the following OLS model:

$$Y_i^{\text{Understanding of LRDG}} = \alpha + \varphi_{\text{Condition}} [\text{label}]_i + \varphi_{\text{individual}} + \beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + e_i$$

Where:

$Y_i^{\text{Awareness/Understanding of LRDG}}$  was the outcome measure for each individual  $i$ .

$\alpha$  was the constant.

$\text{Condition} [\text{label}]_i$  was a vector of binary variables indicating which of the 7 label types the participant was randomly assigned to see.

$\varphi_{\text{individual}}$  was a vector of binary variables indicating each participants' (1) age category, (2) social grade, (3) ethnicity, (4) education, and (5) region, coded as described in Table 3.

*Gender* and *Audit C score* were binary and continuous variables respectively and were coded in the manner described in Table 3.

$e_i$  was the error term.

### Secondary Analysis I (secondary outcomes)

4. **Perceived risk.** Participants were asked "To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease?"
5. **Motivation to drink.** Participants were asked: "Earlier, you saw the following alcohol label: To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less."
6. **Subjective perception of 'high risk' drinking.** Participants were asked "Lastly, we are interested in your opinion about the following: How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health?"

Outcomes 4 and 5 were coded on a 1-5 scale (strongly disagree - strongly agree). Outcome 6 was coded as a continuous variable.

All secondary outcomes were examined using the following OLS model:

$$Y_i^{\text{Perceived risk / motivation / high-risk perception}} = \alpha + \varphi_{\text{Condition}_i} + \varphi_{\text{Demographics}_i} + \beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + \beta_3 \text{Smoking}_i + e_i$$

Where:

$Y_i^{\text{Perceived risk / motivation / high-risk perception}}$  was the outcome measure for each individual  $i$ .

$\alpha$  was the constant.

$\text{Condition}_i$  was a vector of binary variables indicating which of the 14 conditions (7 labels \* 2 warnings) the participant was randomly assigned to.

$\varphi_{\text{Demographics}_i}$  was a vector of binary variables indicating each participants  $i$ 's (1) age category, (2) social grade, (3) ethnicity, (4) education, and (5) region, coded as described in Table 3.

$\text{Gender}$ ,  $\text{Smoking}$  and  $\text{Audit C score}$  were two binary and one continuous variable respectively and were coded in the manner described in Table 3.

$e_i$  is the error term.

### Secondary Analysis II (subgroup analysis)

Finally, we repeated the above regressions for all outcomes, but this time while interacting treatment assignment with several personal characteristics (age, gender, Audit C score) in order to understand whether the effectiveness of the intervention materials differed systematically depending on these characteristics (e.g. whether they worked well for older people but not younger people, for heavy drinkers but not light drinkers, etc).

## 1.9 Power Calculations

BIT runs power calculations for every trial to assess whether we can be sufficiently confident that we can detect a difference between the intervention and the control material. This is based on the number of individuals participating in each of the test conditions, the variance in responses, and insights from academic literature and previous studies on the impact of the intervention tested.

In our power calculations, we follow current best practice<sup>12</sup> by adopting a baseline significance threshold for the p-value of our statistical tests of 5%. In addition, we aim to have sufficient statistical power to detect an effect, should it exist, with 80% confidence.

This trial was powered with respect to the first primary outcome (awareness of the LRDG). Recent surveys of nationally-representative samples of the UK population have found that 8-19% of people are aware of this guideline.<sup>13,14</sup> We therefore estimated that 13% of participants in the control group in our experiment would correctly identify the LRDG as '14 units per week'.

Table 4 shows the results of our power analysis. Given the large number of comparisons made in our analysis (i.e. we examine variation in the 3 primary outcomes across 7 types of alcohol label), our power analysis corrected for multiple comparisons in line with standard BIT procedure (described in Appendix Section 4).

We find that we were powered to detect increases the proportion of participants correcting

<sup>12</sup> List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4), 439.

<sup>13</sup> Rosenberg, G., Bauld, L., Hooper, L., Buykx, P., Holmes, J., Vohra, J. 2018. New national alcohol guidelines in the UK: public awareness, understanding and behavioural intentions. *Journal of Public Health*.

<sup>14</sup> Alcohol Health Alliance UK. [How we drink, what we think. Public views on alcohol and alcohol policies in the UK](#). 2018.

identifying the LRDG by 35-49% in relative terms; we considered this realistic given that many of the labels tested explicitly told participants what the LRDG was.

**Table 4: Power analysis results.**

Total N	# arms	N per arm	p-value significance threshold	% control	% treatment	Effect size in percentage points (%)
7000	7	1000	0.002380952 [0.05*(1/21)] [Most strict]	13%	19.4%	6.4 pp (49% increase)
7000	7	1000	0.026190476 [0.05*(11/21)] [Mid-point]	13%	18.0%	5 pp (38% increase)
7000	7	1000	0.05 [0.05*(21/21)] [Least strict]	13%	17.5%	4.5 pp (35% increase)

*Rows 1-3 in Column 4 impose decreasingly strict statistical significance thresholds, mimicking the Hochberg correction procedure employed in the analysis. Note that '21' is the number of comparisons being made (7 label types across 3 primary outcomes).*

*Outcome measure = Binary indicator of whether participants identify '14 units' as the low risk drinking guideline (coded as 1) vs any other response (coded as 0).*

Accepte

identifying the LRDG by 35-49% in relative terms; we considered this realistic given that many of the labels tested explicitly told participants what the LRDG was.

**Table 4: Power analysis results.**

Total N	# arms	N per arm	p-value significance threshold	% control	% treatment	Effect size in percentage points (%)
7000	7	1000	0.002380952 [0.05*(1/21)] [Most strict]	13%	19.4%	6.4 pp (49% increase)
7000	7	1000	0.026190476 [0.05*(11/21)] [Mid-point]	13%	18.0%	5 pp (38% increase)
7000	7	1000	0.05 [0.05*(21/21)] [Least strict]	13%	17.5%	4.5 pp (35% increase)

Rows 1-3 in Column 4 impose decreasingly strict statistical significance thresholds, mimicking the Hochberg correction procedure employed in the analysis. Note that '21' is the number of comparisons being made (7 label types across 3 primary outcomes).

Outcome measure = Binary indicator of whether participants identify '14 units' as the low risk drinking guideline (coded as 1) vs any other response (coded as 0).

Accepte

## 1.10 Risks

Finally, Table 5 describes the two main risks involved with this trial.

**Table 5: Power analysis results.**

Risk	Strategy to mitigate risk	Responsibility	Timeframe (if applicable)
Recruitment of target sample is slower than expected, which can extend time needed for data collection	Increasing amount of financial compensation to take part. Relaxing screening criteria (e.g. do not require strictly nationally-representative age or geographic coverage of England).	BIT	Monitor traffic while the experiment is live and make a decision on launching boosters by the end of day 5.
Poor quality responses from participants due to fatigue or general lack of attention	Adequately compensate respondents & keep length of the experiment to 8-10 minutes.	BIT	Trial design phase

Accepted

## Part 2

### 2.1 Implementation

In January 2019, we ran four pilot studies (total N = 644) to test the presentation of the experimental material. For each pilot iteration, we made small tweaks to the presentation of the experimental material to ensure participants understood what they were being asked to do.

The results of the pilot are summarised in Appendix 5. Based on those results, we made the following changes to the material before launching the main experiment.

**1. We abandoned the planned approach of testing warnings within-arms in a 50-50 ratio.**

We ran the main experiment such that 7000 people were randomly assigned to one of the seven arms (i.e. 1000 per arm) - they saw only an alcohol label *without* a warning.

We also admitted an additional 500 people to the experiment; this group was randomly assigned across the seven arms (i.e. 71 per arm) - they saw an alcohol label *with* a warning. We achieved this by randomly assigning 7% of the sample to see warnings.

Our primary analysis was restricted to the first group of 7000 people. Our secondary analysis compared the responses of the two groups (i.e. the group of 7000 vs the group of 500).

**2. We tweaked the presentation of the 'Understanding' questions (described on p16-17 of this TP).**

(i) For **Spirits III (container) & Spirits IV (container)**, we originally asked "*How much of a bottle or how many bottles (700ml/1L) could you have before reaching 14 units?*"

This was changed to "*How many bottles (700ml/1L) could you have before reaching 14 units?*". We also added the sentence "*You can answer in terms of fractions (e.g. 0.5, 1.3) or whole numbers (e.g. 2, 3)*".

(ii) For **Spirits I (serving) & Spirits II (serving)**, we changed the question from "*How many measures (25ml) could you have...*" to "*How many shots (25ml) could you have...*".

(iii) For **Wine I (serving), Wine II (serving), Spirit I (serving), & Spirit II (serving)**, we added the text "*Your answer must be at least 1*" for people who attempted to enter values

lower than this.

(iv) For **Wine III (container)** & **Wine IV (container)**, we added the text " *Your answer must be at most 10*" for people who attempted to enter values higher than this.

(v) For **Spirits III (container)** & **Spirits IV (container)**, we added the text " *Your answer must be at most 5*" for people who attempted to enter values higher than this.

### 3. We altered our secondary analysis strategy so that we no longer examined responses across all 14 possible cells (ie 7 labels \* 2 warning conditions)

Our secondary analysis strategy is now the below:

#### REVISED Secondary Analysis strategy

**Perceived risk.** Participants were asked " *To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease?*"

**Motivation to drink.** Participants were asked: " *Earlier, you saw the following alcohol label: To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less.*"

**Subjective perception of 'high risk' drinking.** Participants were asked " *Lastly, we are interested in your opinion about the following: How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health?*"

Outcomes 4 and 5 were coded on a 1-5 scale (strongly disagree - strongly agree).  
Outcome 6 was coded as a continuous variable.

We examined the three secondary outcomes using the following OLS models:

$$Y_{it} = \alpha + \varphi_{it} \text{Condition}_{[label]i} + \varphi_{it} \text{Control Variables} + \beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + \beta_3 \text{Smoking}_i + e_i$$

$$Y_{it} = \alpha + \varphi_{it} \text{Condition}_{[warning]i} + \varphi_{it} \text{Control Variables} + e_i$$



$$\beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + \beta_3 \text{Smoking}_i + e_i$$

Where:

$y_i^{\text{Perceived risk / motivation / high-risk perception}}$  was the outcome measure for each individual  $i$ .

$\alpha$  was the constant.

$\text{Condition}_{[\text{label}]i}$  was a vector of binary variables indicating which of the 7 labels the participant was randomly assigned to.

$\text{Condition}_{[\text{warning}]i}$  was a binary variable indicating which of the 2 warning conditions (warning or no warning) the participant was randomly assigned to.

$\varphi_{\text{Demographic}}^i$  was a vector of binary variables indicating each participants  $i$ 's (1) age category, (2) social grade, (3) ethnicity, (4) education, and (5) region, coded as described in Table 3.

*Gender*, *Smoking* and *Audit C score* were two binary and one continuous variable respectively and were coded in the manner described in Table 3.

$e_i$  is the error term.

We also added this additional secondary analysis:

$$y_i^{\text{Awareness of LRDG}} = \alpha + \varphi_{\text{Condition}}^i \text{Condition}_{[\text{warning}]i} + \varphi_{\text{Demographic}}^i + \beta_1 \text{Gender}_i + \beta_2 \text{Audit C score}_i + \beta_3 \text{Smoking}_i + e_i$$

#### 4. We included an attention check question in the main experiment

In the pilot, we included an attention check question towards the end of the experiment. This asked "his question is to check whether you are paying attention. Please choose the answer 'Agree'". Participants were given five answer options (ranging from 'Strongly Disagree' to 'Strongly Agree') - we found that 94% of them chose the 'Agree' answer, indicating they were paying attention. Comparing the response patterns of people who passed vs failed the attention check revealed that people who failed:

- completed the survey much quicker (median completion time of 4m32s vs 7m16s for people who passed the attention check), and
- were much less likely to answer the 'Awareness' question correctly (0% answered correctly vs 33% for people who passed the attention check).

Based on these findings, we decided to carry over the attention check question into the main experiment, and to exclude people who failed the check from the final analysis.

## 2.2 Results

### 2.2.1 Balance checks

	Control	T1	T2	T3	T4	T5	T6
Female	49.4%	51.9%	50.8%	49.8%	52.6%	49.5%	49.5%
p-val		0.25	0.52	0.86	0.14	0.97	0.97
Age (avg)	44.2	43.6	44.0	43.9	43.9	44.1	45.3
p-val		0.42	0.85	0.74	0.73	0.91	0.13
Social grade (avg) [1 = highest, 6 = lowest]	3.7	3.7	3.7	3.7	3.8	3.7	3.8
p-val		0.97	0.68	0.87	0.17	0.76	0.35

Green = Characteristic is not significantly different in treatment groups vs control (ie  $p > 0.05$ ).

Red = Characteristic is significantly different in treatment groups vs control (ie  $p < 0.05$ ).

## 2.2.2 Descriptives

Table S1. Full characteristics of the sample (N=7521)

Characteristic	% of sample		Characteristic	% of sample
<b>Gender</b>			<b>Social grade</b>	
Female	50.5%		A	7.5%
Male	49.5%		B	24.6%
			C1	24.4%
<b>Age</b>			C2	1.8%
18-24	13.4%		D	15.6%
25-54	52.3%		E	26.1%
55+	34.3%			
			<b>Ethnicity</b>	
<b>Region</b>			White	90.9%
North	28.9%		Black	2.5%
South & East	36.3%		Asian	3.7%
Midlands	20.3%		Other	2.9%
London	14.6%			
			<b>Audit C score</b>	
<b>Income</b>			1-3	32.8%

Less than £30k	52.4%		4-6	38.7%
£30k+	47.6%		7-9	20.7%
			10-12	7.8%
Smoker				
No	72.2%		Highest education	
Yes	27.8%		None	1.9%
Saw alcohol label with warning			Secondary	23.2%
No	93.4%		Post-secondary / Vocational	37.7%
Yes	6.6%		Undergrad or higher	37.2%

### 2.2.3 Primary analysis

Table S2. Awareness of the LRDG, regression analysis (N=7521)

	Whether participants said the LRDG was 14 units per week (0 = no, 1 = yes)		Whether participants said the LRDG was 14 units per week (0 = no, 1 = yes)	
Sample size	7521		7521	
Characteristic	Logit marginal effect	p-value	Odds ratios [95% CIs]	p-value

	95% CIa]			
<b>Treatment (base = control)</b>				
Food label (serving)	17.3*** (13.6, 21.0)	<0.001	2.4*** (2.0, 3.0)	<0.001
Food label (serving & container)	11.4*** (7.9, 15.0)	<0.001	1.9*** (1.5, 2.3)	<0.001
Pictograph (serving)	27.0*** (23.3, 30.7)	<0.001	3.7*** (3.1, 4.5)	<0.001
Pictograph (container)	29.3*** (25.5, 33.0)	<0.001	4.1*** (3.4, 5.0)	<0.001
Pie chart	25.9*** (22.0, 29.7)	<0.001	3.5*** (2.9, 4.3)	<0.001
Risk gradient	13.9*** (10.2, 17.6)	<0.001	2.1*** (1.7, 2.6)	<0.001
<b>Age (base = 18-24)</b>				
25-54	7.4*** (4.4, 10.5)	<0.001	1.5*** (1.2, 1.7)	<0.001
55+	26.0*** (22.7, 29.4)	<0.001	3.3*** (2.8, 3.9)	<0.001

Female (vs male)	6.9*** (4.8, 9.0)	<0.001		1.4*** (1.2, 1.5)	<0.001
Social grade C2DE (vs ABC1)	-2.7* (-4.9, -0.4)	0.02		0.9* (0.8, 1.0)	0.02
Ethnicity (base = white)					
Black	-3.9 (-11.6, 3.8)	0.32		0.8 (0.6, 1.2)	0.33
Asian	-9.5** (-15.3, -3.7)	<0.01		0.6** (0.5, 0.8)	<0.01
Mixed	-11.3** (-18.3, -4.3)	<0.01		0.6** (0.4, 0.8)	<0.01
Other	-3.1 (-16.0, 9.9)	0.64		0.9 (0.5, 1.6)	0.65
Region (base = North)					
South & East	0.8 (-1.8, 3.5)	0.53		1.0 (0.9, 1.2)	0.53
Midlands	1.0 (-2.0, 4.1)	0.51		1.0 (0.9, 1.2)	0.51
London	-6.4*** (-9.8, -2.9)	<0.001		0.7*** (0.6, 0.9)	<0.001

Audit C (1-12)	1.5*** (1.1, 1.9)	<0.001		1.1*** (1.1, 1.1)	<0.001
<b>Education (base = none)</b>					
Secondary	13.1*** (5.9, 20.2)	<0.001		2.0*** (1.3, 3.2)	<0.001
Post-secondary / Vocational	19.3*** (12.3, 26.3)	<0.001		2.8*** (1.8, 4.3)	<0.001
Undergrad or higher	20.6*** (13.5, 27.7)	<0.001		2.9*** (1.9, 4.6)	<0.001
Warning (vs no warning)	-0.6 (-4.9, 3.6)	0.77		1.0 (0.8, 1.2)	0.77

Constant is not produced in marginal effect model, and has been omitted from presentation of odds ratio model. \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$

Table S3. Understanding of the LRDG

Characteristic	Servings: Distance to correct answer			Containers: Distance to correct answer	
	OLS coefficient [95% CIe]	p-value		OLS coefficient [95% CIe]	p-value
Sample size	7483			7505	
<b>Treatment (base = control)</b>					
Food label (serving)	3.4*** (3.2, 3.7)	<0.001		1.0*** (0.9, 1.1)	<0.001



Food label (serving & container)	3.2*** (3.0, 3.5)	<0.001		0.3*** (0.2, 0.4)	<0.001
Pictograph (serving)	3.7*** (3.4, 3.9)	<0.001		0.8*** (0.7, 0.9)	<0.001
Pictograph (container)	1.2*** (0.9, 1.4)	<0.001		0.1** (0.0, 0.2)	<0.01
Pie chart	3.5*** (3.3, 3.8)	<0.001		0.7*** (0.6, 0.8)	<0.001
Risk gradient	2.8*** (2.5, 3.1)	<0.001		0.7*** (0.6, 0.8)	<0.001
<b>Age (base = 18-24)</b>					
25-54	-0.1 (-0.3, 0.1)	0.42		-0.3*** (-0.4, -0.2)	<0.001
55+	0.4*** (0.2, 0.6)	<0.001		-0.3*** (-0.4, -0.2)	<0.001
<b>Female (vs male)</b>	0.0 (-0.1, 0.2)	0.78		-0.1*** (-0.1, -0.0)	<0.001
<b>Social grade C2DE (vs ABC1)</b>	-0.3*** (-0.4, -0.1)	<0.001		-0.0 (-0.1, 0.0)	0.48
<b>Ethnicity (base = white)</b>					

Black	-0.6** (-1.1, -0.2)	<0.01		0.2* (0.0, 0.4)	0.02
Asian	-0.6** (-0.9, -0.2)	<0.01		0.3*** (0.2, 0.5)	<0.01
Mixed	-0.6** (-1.1, -0.2)	<0.01		0.4*** (0.2, 0.6)	<0.01
Other	-0.1 (-0.9, 0.7)	0.78		0.4* (0.0, 0.7)	0.03
<b>Region (base = North)</b>					
South & East	0.2* (0.0, 0.4)	0.05		-0.0 (-0.1, 0.0)	0.33
Midlands	-0.2 (-0.4, 0.0)	0.11		-0.1 (-0.1, 0.0)	0.08
London	-0.4*** (-0.6, -0.2)	<0.001		0.0 (-0.0, 0.1)	0.26
<b>Audit C (1-12)</b>	0.0 (0.0, 0.0)	<0.001		0.0 (-0.0, 0.0)	0.97
<b>Education (base = none)</b>					
Secondary	0.7* (0.2, 1.2)	<0.05		-0.1 (-0.3, 0.1)	0.60
Post-secondary Vocational	1.2*** (0.7, 1.7)	<0.001		-0.1 (-0.3, 0.1)	0.26

Undergrad or higher	1.3*** (0.8, 1.8)	<0.001		-0.2* (-0.4, 0.0)	0.04
Warning (vs no warning)	0.0 (-0.2, 0.3)	0.74		0.0 (-0.1, 0.2)	0.41
Constant	-5.7*** (-6.3, -5.1))	<0.001		0.5*** (0.3, 0.8)	<0.001
R-squared	0.18			0.10	

The 'Servings' outcome was measured by taking the average of people's estimates for how many beers (2 questions), servings of wines (2 questions), and servings of spirits (2 questions) it takes to reach 14 units, and then subtracting the technically correct answer from this. The analysis excludes 38 participants who gave ineligible responses for at least one of these 6 questions. The 'Containers' outcome was measured by taking the average of people's estimates for how many beers (2 questions), containers of wines (2 questions), and containers of spirits (2 questions) it takes to reach 14 units, and then subtracting the technically correct answer from this. The analysis excludes 16 participants who gave ineligible responses for at least one of these 6 questions.

\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$

## Regression results

Figures 7-9 show the results of our primary regressions (statistical tests to assess whether differences in participants' responses were caused by the different alcohol labels they saw, rather than being due to chance). The full regression results are in Tables S2-S4 in Appendix B. *Note that figures and text is copied from the final report. Descriptive figures are not copied over therefore figure numbers start at 7 in this document. Please refer to the final external report to see descriptive figures.*

The main findings are:

1. Participants in all 6 treatment groups were significantly more likely ( $p < 0.001$ ) than those in the Control to say the LRDG was '14' units per week.[1] This is not very surprising given that the Control label did not explicitly tell people the 14 unit per week guideline, whereas all the treatment labels did.

2. **Understanding of how many servings (of beers, wines, and spirits) they could have before reaching 14 units was most accurate in the Pictograph (Serving) group and least accurate in the Control Group.** Control participants estimated they could have 4.6 fewer servings than the LRDG actually allows, and Pictograph (Serving) participants thought they could have 0.9 fewer servings than actually allowed. All 7 groups underestimated how many servings they could have.
3. **Understanding of how many containers (of beers, wines, and spirits) they could have before reaching 14 units was most accurate in the Control group, and least accurate in the Food Label (Serving) group.** Control participants estimated they could have 0.1 more containers than actually allowed, and participants in the Food Label (Serving thought) they could have 1.1 more containers than actually allowed. All 7 groups overestimated how many containers they could have.
4. **People's inaccurate estimates of how many servings and containers they could have before reaching 14 units were driven almost entirely by their estimates for wine and (especially) spirits. In contrast, participants in all 7 groups gave strikingly accurate estimates of how many beers they could have before reaching 14 units.** For example, Figures 8a and 8b show that while Control participants estimated they could have 4.6 fewer servings (averaged across beer, wine, and spirits) than actually allowed, this estimate breaks down into 0.5 fewer beers, 2.4 fewer glasses of wine, and 11 fewer shots of spirits. Similarly, Figures 9a and 9b show that while Food Label (Serving) participants estimated they could have 1.1 more containers (averaged across beer, wine, and spirits) than actually allowed, this breaks down into 0.4 fewer beers, 2.8 more containers of wine, and 0.9 containers of spirits.
5. **The extent of people's overestimates of how many containers of wine and (especially) spirits they could have before reaching the LRDG was dramatic.** For example, Figure 9c shows that Food Label (Serving) participants overestimate that they can have 0.9 more containers of spirits than is allowed by the LRDG translates into an extra 33.6 units above the 14 unit guideline. Figure 8c shows the opposite problem occurs when people estimate in servings. For example, the Pictograph (Container) group's underestimate that they could have 1.4 fewer servings of wine than is actually allowed by the LRDG translates into 2.9 units below the 14 unit guideline. As evident in these two examples, This indicates that the problem is lopsided - when people overestimate how many containers they can have, this translates into an overestimate of many more units than allowed by the LRDG, but when they underestimate how many servings they can have, this translates into a relatively modest underestimate of how many units they can have.

**Figure 7. Awareness of the LRDG, by treatment group**

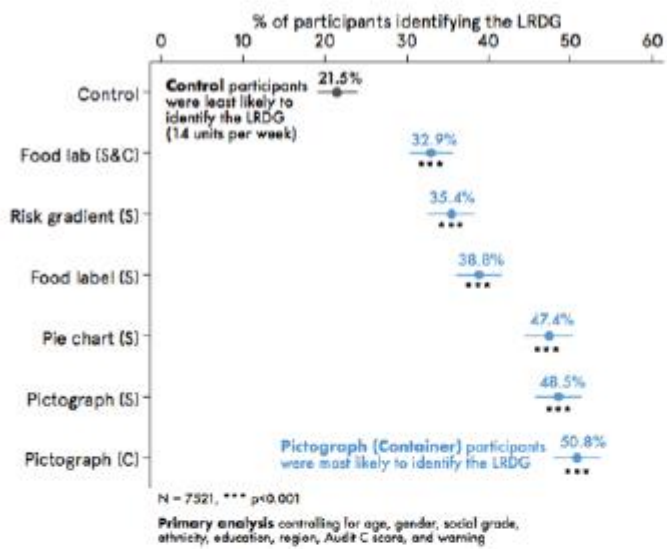


Figure 8a. Understanding (serving) of the LRDG, by treatment group

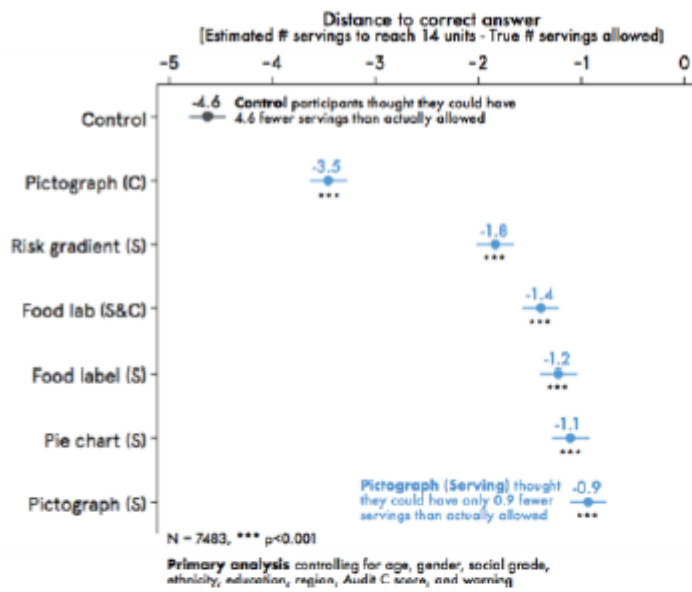


Figure 8b. Understanding (serving) of the LRDG, by treatment group and disaggregated by alcohol type

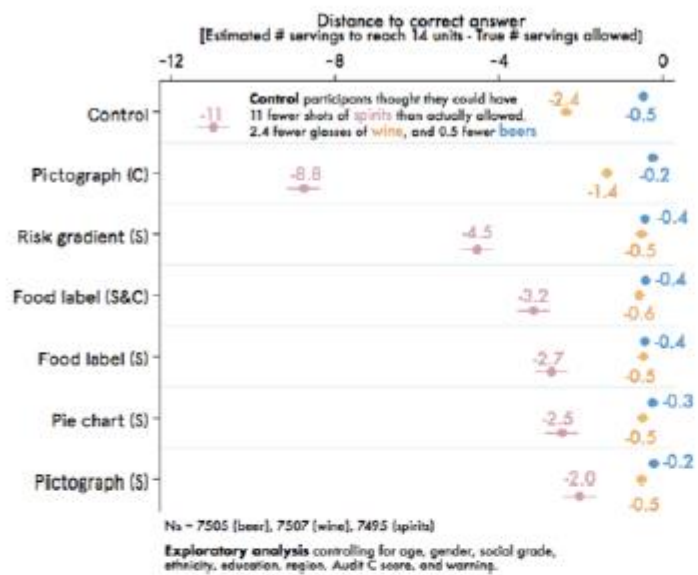


Figure 8c. Understanding (serving) of the LRDG, by treatment group, disaggregated by alcohol type, and with participant estimates converted into units

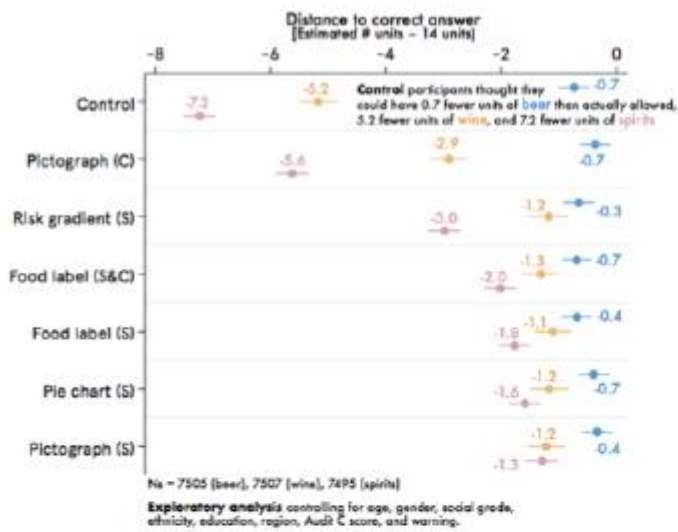


Figure 9a. Understanding (container) of the LRDG, by treatment group

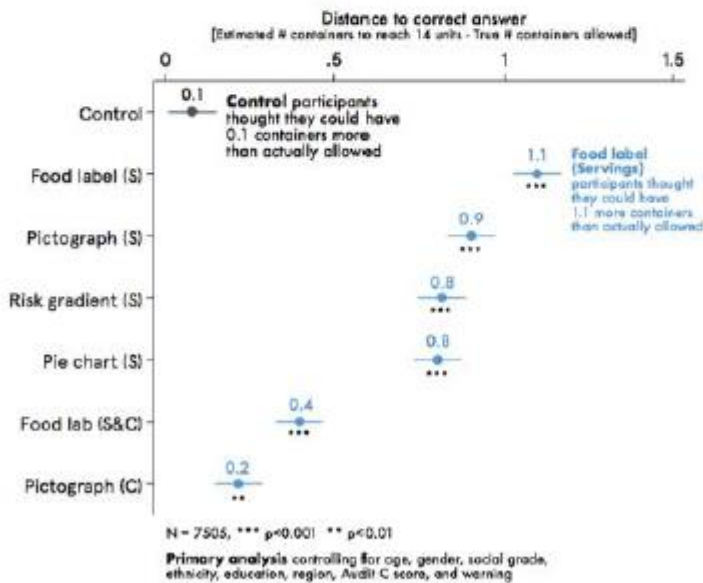




Figure 9b. Understanding (container) of the LRDG, by treatment group and disaggregated by alcohol type

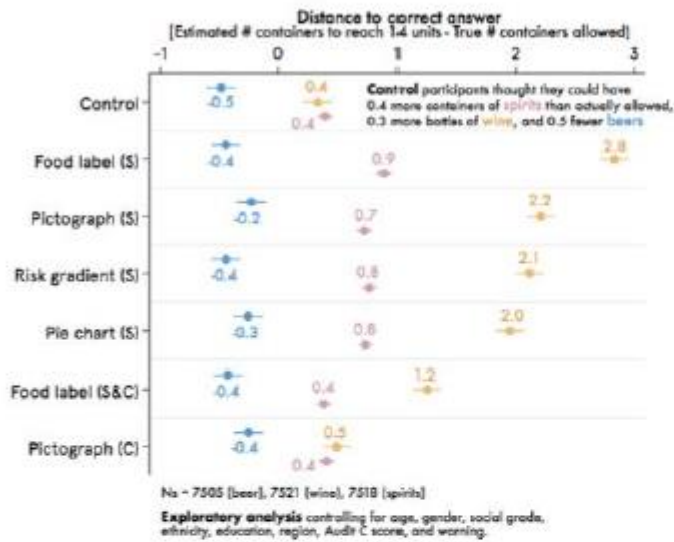
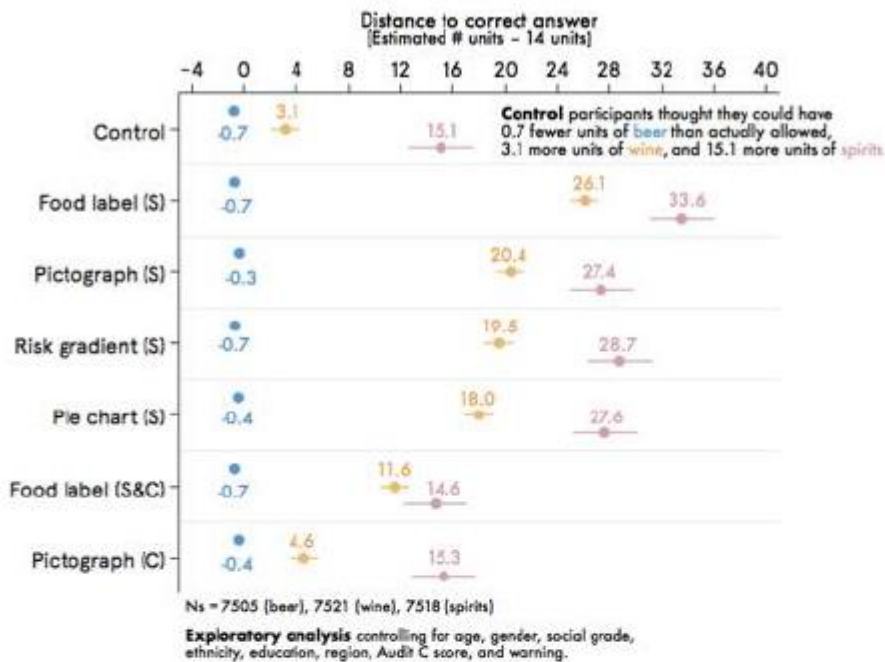


Figure 9c. Understanding (container) of the LRDG, by treatment group, disaggregated by alcohol type, and with participant estimates converted into units



[1] Each of our 3 primary outcomes involved comparing participant responses in 6 treatment groups against the responses of people in a Control group, meaning we made a total of 18 statistical comparisons. To account for this large number of comparisons, we prespecified in an internal Trial Protocol that we would conduct a multiple comparisons corrections procedure to minimize the risk of identifying spurious correlations. Given the highly significant nature of the treatment effects for all three primary outcomes (e.g. of the 18 treatment coefficients, 17 were significant at  $p < 0.001$  and the other was significant at  $p < 0.01$ ), we found that the application of this procedure did not meaningfully alter any of our findings.

## 2.2.4 Secondary analysis

Table S4. Perceived risk of alcohol and motivation to drink

Characteristic	1: Perceived personal risk of own drinking		2: Less motivated to drink		3: Subjective perception of high risk drinking	
	Coefficient [95% CI]	p-value	Coefficient [95% CI]	p-value	Coefficient [95% CI]	p-value
<b>Sample size</b>	7521		7521		7521	
<b>treatment (base = control)</b>						
Food label (serving)	-0.0 (-0.1, 0.1)	0.44	0.1** (0.1, 0.2)	<0.01	0.8 (-0.7, 2.2)	0.30
Food label (serving container)	0.0 (-0.1, 0.1)	0.64	0.2*** (0.1, 0.3)	<0.001	0.6 (-0.8, 2.0)	0.41
Pictograph (serving)	0.0 (-0.1, 0.1)	0.64	0.2*** (0.1, 0.3)	<0.001	0.9 (-0.5, 2.3)	0.22
Pictograph (container)	-0.0 (-0.1, 0.1)	0.89	0.3*** (0.2, 0.4)	<0.001	0.2 (-1.3, 1.6)	0.83
Pie chart	-0.0 (-0.1, 0.1)	0.56	0.2*** (0.1, 0.3)	<0.001	1.2 (-0.2, 2.6)	0.11
Risk gradient	-0.0 (-0.1, 0.1)	0.31	0.2*** (0.1, 0.3)	<0.001	-0.1 (-1.5, 1.4)	0.94
<b>Age (base = 18-44)</b>						
5-54	-0.0 (-0.1, 0.1)	0.68	0.0 (-0.1, 0.1)	0.99	1.6** (0.4, 2.7)	<0.01

5+	-0.1** (-0.2, -0.0)	<0.01	-0.1* (-0.2, -0.0)	0.02	5.1*** (3.8, 6.3)	<0.001
Female (vs male)	0.1*** (0.0, 0.2)	<0.001	0.0 (-0.0, 0.1)	0.08	-0.3 (-1.1, 0.5)	0.47
Social grade (2DE vs ABC1)	-0.0 (-0.1, 0.0)	0.39	-0.0 (-0.1, 0.0)	0.61	0.7 (-0.2, 1.5)	0.12
Ethnicity (base = white)						
Black	-0.1 (-0.3, 0.1)	0.29	0.1 (-0.1, 0.2)	0.53	0.0 (-2.6, 2.7)	0.98
Asian	0.1 (-0.0, 0.2)	0.12	0.2*** (0.1, 0.4)	<0.001	-2.5* (-4.5, -0.4)	0.02
Mixed	0.2* (0.0, 0.3)	0.04	0.2** (0.1, 0.4)	<0.01	0.8 (-1.7, 3.3)	0.52
Other	-0.1 (-0.4, 0.2)	0.72	0.1 (-0.1, 0.3)	0.32	1.1 (-3.4, 5.5)	0.63
Region (base = North)						
South & East	-0.0 (-0.1, 0.0)	0.63	-0.0 (-0.1, 0.0)	0.10	0.7 (-0.3, 1.6)	0.17

Midlands	0.0 (-0.1, 0.1)	0.63	-0.0 (-0.1, 0.1)	0.94	0.7 (-0.4, 1.8)	0.23
London	-0.0 (-0.1, 0.1)	0.86	0.1* (0.0, 0.2)	0.04	-1.0 (-2.3, 0.3)	0.12
Audit C (1-12)	-0.0 (-0.0, 0.0)	0.65	-0.0*** (-0.1, -0.0)	<0.001	1.6*** (1.4, 1.7)	<0.001
<b>Education (base = none)</b>						
Secondary	0.3** (0.1, 0.5)	<0.01	0.0 (-0.1, 0.2)	0.72	3.5* (0.6, 6.3)	0.02
Post-secondary / Vocational	0.3** (0.1, 0.5)	<0.01	-0.0 (-0.2, 0.2)	0.9167	2.7 (-0.1, 5.5)	0.06
Undergrad or higher	0.3** (0.1, 0.5)	<0.01	0.0 (-0.1, 0.2)	0.	2.2 (-0.7, 5.1)	0.13
Warning (vs no warning)	-0.0 (-0.1, 0.1)	0.94	0.1 (-0.0, 0.2)	0.12	-0.3 (-1.9, 1.2)	0.67
Constant	3.6*** (3.4, 3.8)	<0.001	3.3*** (3.1, 3.5)	<0.001	10.4*** (7.1, 13.8)	<0.001
R squared	0.01		0.03		0.08	

'Perceived personal risk of drinking' was measured by asking "To what extent do you think that cutting down on your drinking would reduce your own risk of alcohol related disease?"

'Motivation to drink' was measured by asking "Earlier, you saw the following alcohol label: [beer image #3]. To what extent do you agree or disagree with the following statement: This information makes me feel motivated to drink less."

*'Subjective perception of high risk drinking' was measured by asking "How many units of alcohol do you personally think a person would need to regularly drink per week to seriously damage their health?"*

*\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$*

### 2.2.5 Secondary analysis II - subgroup analysis

The final portion of our analysis examined whether the 7 alcohol labels (and cancer warning) had different effects on the primary and secondary outcomes depending on 5 participant characteristics: gender, age, social grade, ethnicity, and Audit C score.

The motivation for this was to check whether the intervention materials exacerbated health inequalities. For example, the preceding section showed that the treatment alcohol labels in this experiment increased awareness of the LRDG. However, we would be concerned if this was achieved this by, for example, only improving awareness among university educated participants (thereby increasing the awareness gap between the most and least educated members of society).

In brief, we found that the vast majority of our subgroup analyses - shown in Appendix C - identified no significant effects (i.e.  $p > 0.05$ ).<sup>[1]</sup> In other words, we did not find strong evidence that the different alcohol labels (and cancer warning) notably increased inequality (in terms of knowledge about alcohol risk) between men and women, old and young, whites and people of other ethnicities, higher (ABC1) and lower (C2DE) social grades, or between lighter and heavier drinkers.

There were a small number of exceptions, which we show graphically in Appendix C. For example, among people who saw the Control alcohol label, we found that 21% of men and 23% of women correctly identified the '14 unit' LRDG (i.e. there was a 2 percentage point gender gap). However, among people who saw the risk gradient design, 31% of men and 40% of women answered this correctly (i.e. there was a 9 percentage point gender gap). Compared to the Control label, the risk gradient design therefore increased the knowledge gap of the LRDG by 7 percentage points, and this increase was statistically significant (at  $p < 0.05$ ). However, we note that given the large number of subgroup analyses conducted, it is possible that these differences are due to chance, even though they passed the conventional threshold of statistical significance ( $p < 0.05$ ).

---

[1] Although we conducted a large amount of subgroup analysis, we did not apply a multiple comparisons corrections procedure to the results. BIT's standard practice is to apply these procedures only to primary analysis, and not to secondary and exploratory analysis (respectively, analyses which were and were not prespecified in an internal Trial Protocol) because we already consider these latter analyses to be more speculative.

**C1. Whether the effect of the alcohol labels on the primary and secondary outcomes differed depending on 5 demographic characteristics**

**Table S5. Summary of whether, compared to the Control, the treatment labels differently affected certain demographics in terms of Awareness of the LRDG (grey = no effect, red = increased knowledge inequality within a demographic category, green = decreased knowledge inequality within a demographic category).**

Primary outcome: Awareness of the '14 unit' LRDG						
	Alcohol label					
	1	2	3	4	5	6
By age (old vs young)						Green
By gender (women vs men)						Red
By Audit C score (high vs low)						Grey
By ethnicity (white vs other ethnicities)						Grey
By social grade (ABC1 vs C2DE)				Red		Grey

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

**Figure S1. Significant interactions between the alcohol labels and certain demographics in terms of affecting Awareness of the LRDG.**

**Table S6. Summary of whether, compared to the Control, the treatment labels differently affected certain demographics in terms of perceived risk of own alcohol consumption (grey = no effect, red = increased differences in perceived personal risk within a demographic category, green = decreased differences in perceived personal risk within a demographic category).**

Secondary outcome 1: Perceived risk of own alcohol consumption						
	Alcohol label					
	1	2	3	4	5	6
By age (old vs young)	Green	Grey	Green	Grey	Grey	Green
By gender (women vs men)	Grey	Grey	Grey	Grey	Grey	Grey
By Audit C score (high vs low)	Grey	Grey	Grey	Grey	Grey	Grey
By ethnicity (white vs other ethnicities)	Grey	Grey	Grey	Grey	Grey	Grey
By social grade (ABC1 vs C2DE)	Grey	Grey	Grey	Grey	Grey	Grey

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .



**Figure S2. Significant interactions between the alcohol labels and age in terms of affecting perceived risk of alcohol consumption.**

**Table S7. Summary of whether, compared to the Control, the treatment labels differently affected certain demographics in terms of motivation to drink (grey = no effect, red = increased differences in motivation to drink within a demographic category, green = decreased differences in motivation to drink within a demographic category).**

Secondary outcome 2: Motivation to drink						
	Alcohol label					
	1	2	3	4	5	6
By age (old vs young)	Grey	Grey	Grey	Grey	Grey	Grey
By gender (women vs men)	Grey	Grey	Grey	Grey	Grey	Grey
By Audit C score (high vs low)	Grey	Grey	Red	Grey	Grey	Grey
By ethnicity (white vs other ethnicities)	Grey	Grey	Grey	Grey	Grey	Grey
By social grade (ABC1 vs C2DE)	Grey	Grey	Grey	Grey	Grey	Grey

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

**Figure S3. Significant interactions between the alcohol labels and Audit C score in terms of affecting motivation to drink.**

**Table S8. Summary of whether, compared to the Control, the treatment labels differently affected certain demographics in terms of perception of 'high risk' drinking (grey = no effect, red = increased differences in perception of high risk within a demographic category, green = decreased differences in perception of high risk within a demographic category).**

Secondary outcome 3: Perception of 'high risk' drinking						
	Alcohol label					
	1	2	3	4	5	6
By age (old vs young)	Grey	Grey	Grey	Grey	Grey	Grey
By gender (women vs men)	Grey	Grey	Grey	Grey	Grey	Grey
By Audit C score (high vs low)	Red	Grey	Grey	Red	Grey	Grey
By ethnicity (white vs other ethnicities)	Grey	Grey	Grey	Grey	Grey	Grey
By social grade (ABC1 vs C2DE)	Grey	Grey	Grey	Grey	Grey	Grey

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

**Figure S4. Significant interactions between the alcohol labels and Audit C score in terms of affecting perception of 'high risk' drinking.**

**C2. Whether the effect of the cancer warnings on the primary and secondary outcomes differed depending on 5 demographic characteristics**

Table S9. Summary of whether, compared to the Control, the cancer warning differently affected certain demographics in terms of Awareness of the LRDG (grey = no effect, red = increased knowledge inequality within a demographic category, green = decreased knowledge inequality within a demographic category).

Primary outcome: Awareness of the '14 unit' LRDG		
	Saw cancer warning?	
	No	Yes
By age (old vs young)		
By gender (women vs men)		
By Audit C score (high vs low)		
By ethnicity (white vs other ethnicities)		
By social grade (ABC1 vs C2DE)		

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

**Table S10. Summary of whether, compared to the Control, the cancer warning differently affected certain demographics in terms of perceived risk of own alcohol consumption (grey = no effect, red = increased differences in perceived personal risk within a demographic category, green = decreased differences in perceived personal risk within a demographic category).**

Secondary outcome 1: Perceived risk of own alcohol consumption		
	Saw cancer warning?	
	No	Yes
By age (old vs young)		
By gender (women vs men)		
By Audit C score (high vs low)		
By ethnicity (white vs other ethnicities)		
By social grade (ABC1 vs C2DE)		

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

**Table S11. Summary of whether, compared to the Control, the cancer warning differently affected certain demographics in terms of motivation to drink (grey = no effect, red = increased differences in motivation to drink within a demographic category, green = decreased differences in motivation to drink within a demographic category).**

Secondary outcome 2: Motivation drink		
	Saw cancer warning?	
	No	Yes
By age (old vs young)		
By gender (women vs men)		
By Audit C score (high vs low)		
By ethnicity (white vs other ethnicities)		
By social grade (ABC1 vs C2DE)		

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

**Table S12. Summary of whether, compared to the Control, the cancer warning differently affected certain demographics in terms of perception of 'high risk' drinking (grey = no effect, red = increased differences in perception of high risk within a demographic category, green = decreased differences in perception of high risk within a demographic category).**

Secondary outcome 3: Perception of 'high risk' drinking		
	Saw cancer warning?	
	No	Yes
By age (old vs young)		

By gender (women vs men)		
By Audit C score (high vs low)		
By ethnicity (white vs other ethnicities)		
By social grade (ABC1 vs C2DE)		

Labels: 1 = Food label (Serving), 2 = Food label (Serving & Container), 3 = Pictograph (Serving), 4 = Pictograph (Container), 5 = Pie Chart, 6 = Risk gradient. Differences were categorised as 'significant' if below  $p < 0.05$ .

## 2.3 Conclusion

### Summary of main findings

We conducted the largest experiment (N=7521) to date on the effectiveness of different alcohol label designs for improving awareness and understanding of alcohol risk (operationalised using the UK government's low risk drinking guideline (LRDG) of 14 units per week).

Our methodology extended the approach used by a 2018 UK study involving 1884 participants.<sup>[1]</sup> We tested 3 alcohol labels from that study (two food label designs and a pie chart) alongside 3 novel labels (two pictographs and a risk gradient design), and compared all of these against a Control group who saw the 'Responsibility Deal' labels currently used across the UK.

Table 4 summarises our main results - in terms of raising both awareness and understanding of the LRDG, the two Pictograph labels performed best.

**Table 4. The 7 alcohol labels, ranked by accuracy on the primary outcomes**

Accuracy	Awareness of LRDG	Understanding of how many <u>servings</u> it takes to reach 14 units	Understanding of how many <u>containers</u> it takes to reach 14 units
Most	Pictograph (C)	Pictograph (S)	Control

	Pictograph (S)	Pie Chart (S)	Pictograph (C)
	Pie Chart (S)	Food label (S)	Food label (S&C)
	Food label (S)	Food label (S&C)	Pie chart (S)
	Risk gradient (S)	Risk gradient (S)	Risk gradient (S)
	Food label (S&C)	Pictograph (C)	Pictograph (S)
Least	Control	Control	Food label (S)

The low level of awareness (21%) of the LRDG in the Control group is unsurprising, given it was the only label that did not explicitly state the LRDG. This figure is also in line with recent UK surveys, where the proportion of participants correctly reporting the LRDG has varied from 8%<sup>[2]</sup> to 16%<sup>[3]</sup> to 25%<sup>[4]</sup>. This figure was 42% for participants across the 6 treatment arms, all of whom saw a label that explicitly stated the 14 unit LRDG.

Our finding that participants in all 7 label groups underestimated how many servings it takes to reach 14 units, with Control participants giving the largest underestimates, replicates the findings from a previous UK study.<sup>[5]</sup> When averaging across participant estimates for how many servings of beer, wine, and spirits they could have before reaching 14 units, Control participants estimated they could have 4.6 fewer servings (4.4 fewer units) than the LRDG actually allows. Participants in the best performing group, Pictograph Serving, thought they could have 0.9 fewer servings (1 fewer unit) than actually allowed.

We also found that participants in all 7 label groups overestimated how many containers they could have before reaching 14 units. When averaging across their estimates for how many containers of beer, wine, and spirits they could have, Control participants estimated they could have 0.1 more containers (6 more units) than actually allowed. All 6 treatment labels backfired and also made people overestimate how many containers they could have. Participants in the worst performing group, Food Label Serving, thought they could have 1.1 more containers (20 more units) than actually allowed.

Breaking down these results by alcohol type revealed that these inaccurate estimates for servings and containers were driven almost entirely by estimates for wine and (especially) spirits. For example, the worst performing group for serving estimates (Control) underestimated beer servings by 0.5 servings, but underestimated wine by almost

5 times as much (2.4 fewer servings) and spirits by 22 times as much (11 fewer servings). The worst performing group for containers (Food Label Serving) underestimated beer containers by 0.4 containers, yet overestimated wine by more than twice as much as they underestimated beer (0.9 more containers) and spirits by seven times as much (2.8 more containers). The tendency for people to give particularly inaccurate estimates for spirits was also found by a recent UK study.<sup>[6]</sup>

Our secondary analysis found that the vast majority of people thought that cutting down on their drinking would reduce their health risk (88%), which did not notably vary across the different alcohol label groups. A large minority of participants (41%) agreed that the alcohol label they saw motivated them to drink less, and this figure was 6 percentage points higher among people who saw the treatment labels with cancer warnings. Interestingly, only 30% of participants classified as heavy drinkers by their Audit C scores reported that they were motivated to drink less, suggesting that understanding of the LRDG may not necessarily lead to behaviour change.

We also found that, after telling all participants that the LRDG was 14 units per week, participants on average said that they thought a person would need to drink 24 units per week to 'seriously damage' their health, and this average did not notably vary depending on what alcohol label they had seen. Lastly, we found that the cancer warning did not significantly affect people's responses to any of the primary or secondary outcomes.

## Recommendations

Based on the findings of this experiment, we make the following recommendations.

Recommendations	
1	<b>Increase awareness of the 14 unit LRDG by using the Pictograph or Pie Chart designs.</b> Awareness in these groups (47-51%), which paired a relatively uncluttered design with text explicitly telling people the 14 unit guideline, was more than twice as high as the Control group (21%), which showed people industry-standard labels.
2	<b>Explain how many servings (not containers) it takes to reach the 14 unit guideline.</b> People reliably overestimate how many containers it takes to reach 14 units and reliably underestimate how many servings it takes.
	If choosing between explaining alcohol units to people in servings or containers, from a public health perspective it is better if participants underestimate (rather than overestimate) how much they can drink to keep health risks low. We therefore recommend contextualising how much alcohol it takes to reach the LRDG in terms of servings.



- 3 Use the 'Pictographs Servings' design to explain servings. The most effective visual design tested in this experiment was the below - a pictograph approach which talks in terms of bottles of beer, glasses of wine, and shots of spirits.

Our recommendations come with one major caveat - it may not be tenable to implement the Pictograph Serving design for spirits in the real world *if a single serving of spirits is defined using the conventional 25ml measure*. For example, the below image shows one potential consequence of following this convention.

It is mathematically true that a person can have 26 standard 25ml measures of Malibu before reaching the 14 unit guideline, but we suspect many people would be surprised by this fact (as suggested by the quote from one experiment participant). One reason for this is that many people cannot accurately perceive the volume of a single 25ml measure. Indeed, recent studies in Scotland<sup>[7]</sup> and England<sup>[8]</sup> have found that when ordinary people are asked to pour out a 'normal serving' of spirits, they on average pour 2 units of alcohol - similar to a 50ml 'double shot' of a 40% spirit (and about 4 times stronger than a single Malibu shot).

If the Pictograph Servings design was rolled out across the UK tomorrow, we consider it a real possibility that people might incorrectly believe that the LRDG allows them to consume a much greater volume of spirits than is actually the case. This could endanger people's health as a result. This issue could potentially be addressed by communicating spirit servings slightly differently to what was tested in this experiment (e.g. by using designs which talked in terms of the 50ml servings people often tend to pour for themselves, rather than the standard 25ml measure), or by using a different design entirely for spirits.

This is an extreme example of a broader issue. In the experiment, people were very accurate at estimating how many beers they could have before reaching 14 units, less accurate for wine, and least accurate for spirits. We suspect this is because people often drink entire bottles (or cans) of beer as a single serving, but they (usually) need to pour out servings of wine and spirits. As noted by a review of public understanding of alcohol units in multiple countries<sup>[9]</sup>, there is wide variation in how much alcohol different people tend to pour in different situations. We therefore suggest that any future implementation of new alcohol label designs should carefully consider how to help people accurately perceive the volume of liquid in a given serving.

Finally, it is important to consider how the label designs would perform in real life. In our online experiment, the labels were presented on a screen, and though the size of the labels on mobile screens were comparable to the size of labels on a bottle, participants could zoom in if they wished. Labels that display a large amount of information (e.g. the Food label – Serving and Container design) or which are wide in design (e.g. the Risk Gradient) may perform worse in real life if they need to be shrunk down to fit on standard alcohol packaging. Furthermore, although the average participant in our experiment spent around 60 seconds reviewing the various example labels, in the real world, we would expect people to pay less attention to this information. These considerations all make it essential to test the efficacy of novel labels in real-world settings before upscaling them more widely.

- 
- [1] Blackwell, A. K., Drax, K., Attwood, A. S., Munafò, M. R., & Maynard, O. M. (2018). Informing drinkers: Can current UK alcohol labels be improved?. *Drug and alcohol dependence*, 192, 163-170.
- [2] Rosenberg, G., Bauld, L., Hooper, L., Buykx, P., Holmes, J., & Vohra, J. (2017). New national alcohol guidelines in the UK: public awareness, understanding and behavioural intentions. *Journal of Public Health*, 1-8.
- [3] Alcohol Health Alliance UK. (2018, January 10). Awareness of drinking guidelines remains low. Retrieved from <http://www.ias.org.uk/default.aspx?page=2824>
- [4] Buykx, P., Li, J., Gavens, L., Hooper, L., Gomes de Matos, E., & Holmes, J. (2018). Self-reported knowledge, correct knowledge and use of UK drinking guidelines among a representative sample of the English population. *Alcohol and alcoholism*, 53(4), 453-460.
- [5] Blackwell, A. K., Drax, K., Attwood, A. S., Munafò, M. R., & Maynard, O. M. (2018). Informing drinkers: Can current UK alcohol labels be improved?. *Drug and alcohol dependence*, 192, 163-170.
- [6] Blackwell, A. K., Drax, K., Attwood, A. S., Munafò, M. R., & Maynard, O. M. (2018). Informing drinkers: Can current UK alcohol labels be improved?. *Drug and alcohol dependence*, 192, 163-170.
- [7] Gill J, Donaghy M. (2004). Variation in the alcohol content of a 'drink' of wine and spirit poured by a sample of the Scottish population. *Health Education Research*.
- [8] Boniface S, Kneale J, Shelton N. (2013). Actual and perceived units of alcohol in a self-defined "usual glass" of alcoholic drinks in England. *Alcoholism: Clinical and Experimental Research*.
- [9] Kerr WC, Stockwell T. (2012). Understanding standard drinks and drinking guidelines. *Drug and alcohol review*.

## Appendix 1: Full participant instructions and treatment materials

This information is [available here](#).

A trial preview link is [here](#).

## Appendix 2: Power Calculation Code

```
power twoprop 0.13, n(2000) power(0.8) alpha(0.002380952) direction(upper)
power twoprop 0.13, n(2000) power(0.8) alpha(0.026190476) direction(upper)
power twoprop 0.13, n(2000) power(0.8) alpha(0.05) direction(upper)
```

## Appendix 3: Cleaning and Analysis Code

## Appendix 4: BIT's procedure for correcting for multiple comparisons

There are a number of different options for correcting for multiple comparisons, but the premise of these is largely the same. When conducting multiple comparison tests, the burden of proof is raised based on the number of tests that are being conducted, such that your burden of proof across the tests remains broadly constant. The simplest example of this is a Bonferroni correction, which mechanically decreases the type 1 error tolerance for each additional test. For example, if the analysis conducted contains 2 tests, the type 1 error tolerance would be  $0.05/2 = 0.025$ . If five tests were conducted, the tolerance falls to  $0.05/5 = 0.01$ . The effect of this on sample size requirements for studies with more tests is not so mechanical, but for an effect of Cohen's  $d = 0.1$ , the sample size requirements for these tests (with 80% power) are; 1570 per arm, 1902 per arm, and 2337 per arm.

### BIT's approach

The approach taken by BIT is to make use of limitation and pre-registration of analysis for the majority of trials, but to use multiple comparison adjustments when large numbers of tests are included within primary analysis. The grid below outlines this policy, and how it varies according to the number of tests being conducted.

Should I use multiple comparison adjustments? Orange = Yes					
		Number of Outcomes			
Number of Trial Arms		1	2	3	4 or more
	1				Orange
	2				Orange
	3			Orange	Orange
	4		Orange	Orange	Orange
	5 or more	Orange	Orange	Orange	Orange

**How to use the Hochberg step-up procedure:**

Suppose you are running  $k$  hypothesis tests (I will take  $k=5$  in this example). Rank the  $p$ -values from smallest to largest and compare them with a sequence increasing uniformly from  $0.05/k$  to  $0.05$  for the 5% significance level, from  $0.01/k$  to  $0.01$  for the 1% significance level and  $0.1/k$  to  $0.1$  for the 10% significance level, respectively.

Example: you run 5 hypothesis tests, which produce  $p$ -values of

- H1: 0.04
- H2: 0.06
- H3: 0.2
- H4: 0.015
- H5: 0.005

We rank these in increasing order:

- H5:0.005
- H4:0.015
- H1:0.04
- H2:0.06
- H3:0.2

And compare these, if we are looking at a 5% significance level, to a sequence uniformly increasing between 0.05/5 and 0.5:

H5 must be  $< 0.01$  to be accepted (which it **is**)

H4 must be  $< 0.02$  to be accepted (which it **is**)

H1 must be  $< 0.03$  to be accepted (which it **isn't**)

Once you find a hypothesis you reject, then you reject all the rest (in this case H2 and H3). If we had used a Bonferroni correction, we would only have accepted H5 as significant. With no multiple testing adjustment, we would have taken H1 (as well as H4 and H5) as significant.

**How the Hochberg step-up procedure was used for the power analysis in this report**

In this trial, we conducted primary analysis across 7 label types and 3 primary outcomes - 21 trial arms in total. We therefore compared the p-value of the treatment effects in our regression analysis against a sequence beginning with a p-value of 0.002380952 (0.05 multiplied by 1/21) and incrementing uniformly to 0.05 (0.05 multiplied by 21/21). This full sequence of Hochberg corrected p-values is [shown here](#).

## Appendix 5: Summary of pilot findings

Note - the below text is taken from an email sent to the client on 24.1.19; the text has not been altered to make it consistent with the overall tone of this report.

1. We can't proceed with testing the warnings using our planned 50-50 within-arm approach - but we have an alternate solution  
People who saw the warning were 21% more likely (at  $p < 0.1$ ) to answer the 'Awareness' question correctly (i.e. to say that '14 units per week' is the LRDG).

The good news is that this means we now have good preliminary evidence that appending "Warning: Alcohol causes cancer" to alcohol labels does make people pay more attention to them, and therefore be more likely to remember the '14 unit' guideline described in the label.

The less good news is that this finding means we cannot proceed with the planned within-arm approach to testing warnings (ie have 50% of people see warnings and 50% of people not see them) - we pre-specified in the TP that we would not do this if the warning turned out to influence people's responses to any of the primary outcomes (which it has).

We would prefer not to disappoint you by abandoning testing of the warning completely, particularly given your helpful input into its design.

Instead, we propose running the main experiment with 7500 people across the 7 arms.  
- 7000 of these people will not see the warning, meaning we can do our primary analysis as planned without people's responses being contaminated by the warning.  
- We will throw in the extra 500 people for free. These people (ie around 71 extra people in each of the 7 arms) will see both the label and warning.  
- At the end, we will examine the effect of the warnings by comparing these two groups (ie the group of 7000 who don't see the warning vs the group of 500 who see the warning).

2. We've made a number of small but important tweaks to the 'Understanding LRDG' questions to ensure people understand what we are asking them. These changes are:  
(i) Change the phrasing of the 'Understanding LRDG' Spirits Container questions from "How much of a bottle or how many bottles..." to "How many bottles..." and added the extra sentence "You can answer in terms of fractions (e.g. 0.5, 1.3) or whole numbers (e.g. 2, 3)".

These questions showed people an image like this:



**1 bottle = 28 units**



The low risk drinking guideline is  
14 units per week = 0.5 bottles

ABV 40% 700ml

and then asked "How much of a bottle or how many bottles could you have before reaching 14 units".

We expected that most people would answer in terms of fractions, e.g. "0.5".

Instead, many people gave implausibly high values, like "500" or "700". We think these people were answering in terms of ML. In hindsight, this was a reasonable way to answer the question.

We solved this problem by changing the question to "How many bottles could you have before reaching 14 units" and adding the sentence "You can answer in terms of fractions (e.g. 0.5, 1.3) or whole numbers (e.g. 2, 3)".

(ii) Tweak the 'Understanding LRDG' Wine + Spirits questions to reduce the risk of people getting confused about what we are asking when we show servings but ask about containers (or vice versa).

Here is a visual explanation of the problem:

## Problem 1 (show servings, ask about containers)

1. SHOW THIS



2. ASK THIS

How many **bottles of this wine (750ml)** could you have before reaching 14 units?

3. PEOPLE SAY

60% say '7 bottles'

## Problem 2 (show containers, ask about servings)

1. SHOW THIS



2. ASK THIS

How many **measures of this drink (25ml)** could you have before reaching 14 units?

3. PEOPLE SAY

47% say between 0 and 1 (e.g. 0.5)

We don't believe that so many people really think that it takes 7 whole bottles of wine, or only 0.5 shots, in order to reach 14 units.

We think this is people getting confused rather than expressing genuine beliefs because this problem did not happen when we showed people labels about servings and asked about servings, or showed labels about containers and asked about containers. But when we showed servings and asked about containers (or showed containers and asked about servings), many people got confused.

To address this, we suggest these changes:



- For the Spirit Serving questions, change the question from "*How many measures (25ml) could you have*" to "*How many shots (25ml) could you have*". This emphasises to people that, even if they saw a label which talked in terms of *bottles*, we want them to answer in terms of shots. The word 'measures' did not seem to help people realise this, whereas the more familiar word 'shots' did.
- For the Wine Serving (*how many glasses of wine...*) and Spirit Serving (*how many shots*) questions, specify that people cannot give an answer lower than 1. We think that people who give lower values than this (e.g. '*it takes 0.3 shots to reach 14 units*') are much more likely to be answering that way because they misunderstand the question, rather than because they are expressing a genuine belief.
- For the Wine Container questions ("*how many bottles of this wine could you have*"), specify that people cannot give an answer higher than 10. We think that people who give higher values than this (e.g. '*it takes 10 bottles of wine to reach 14 units*') are much more likely to be answering that way because they misunderstand the question, rather than because they are expressing a genuine belief.
- For the Spirit Container questions ("*how many bottles of this drink could you have*"), specify that people cannot not give an answer higher than 5. We think that people who give higher values than this (e.g. '*it takes 26 bottles of Malibu to reach 14 units*') are much more likely to be answering that way because they misunderstand the question, rather than because they are expressing a genuine belief.