MANCHESTER 1824
The University of Manchester

Estimating population average treatment effects for food labeling policy: A causal inference modelling approach for generalizing trial findings to target populations in non-nested designs

A thesis submitted to the University of Manchester for the degree of

Doctor of Philosophy

in the Faculty of Humanities

2026

Constanza Avalos

School of Social Science

Department of Social Statistics

Contents

Word count: 37076

List of figures

List of tables

List of publications

The research presented in this thesis has been disseminated in peer-reviewed journals, is currently under review, or has been presented at international conferences. The publications that directly form the core of this thesis are detailed below. In all of the following works, I served as the lead researcher, responsible for the conceptualisation, methodology, formal data analysis, and writing of the manuscripts.

Peer-reviewed Publications

[P1] Avalos, C., Shryane, N., & Wang, Y. (2026). Food label readability and consumption frequency: isolating content-specific effects via a non-equivalent dependent variable design. *Nutrients, 18(197).*
[Accepted]
** This publication forms the basis of Chapter 3.*

[P2] Avalos, C. (2025). Food label granularity and working memory: effects on food choice in a randomized controlled trial. *Journal of Health, Population and Nutrition, 44(375). [Accepted]*
** This publication forms the basis of Chapter 4.*

[P3] Avalos, C. (Submitted). On the selection of covariates for transportability: a doubly robust BART approach for generalizing a food labeling trial and modeling long-term public health impact. *[Under review] Submitted to the Journal of the Royal Statistical Society: Series A (Statistics in Society).*
** This manuscript forms the basis of Chapter 5.*

Conference Presentations

[C1] Avalos, C. (2024, September). The role and interaction of food labelling systems in promoting healthier and sustainable choices among the UK adult population: A study using choice-based conjoint analysis simulations. *Geneva Health Forum, Geneva, Switzerland.*

[C2] Avalos, C. (2023, September). Food labelling. An experiment to assess consumer choices. *RSS International Conference, Harrogate, United Kingdom.*

[C3] Avalos, C. (2022, June). Food label readability and consumption frequency: isolating content-specific effects via a non-equivalent dependent variable design. *University of Manchester, Manchester, United Kingdom.*

Terms and abbreviations

AI Artificial Intelligence

ANOVA Analysis of Variance

ASMD Absolute Standardized Mean Differences

BART Bayesian Additive Regression Trees

BMI Body Mass Index

CI Confidence Interval

d' d-prime

DAG Directed Acyclic Graph

DiD Difference-in-Differences

DR Doubly Robust

ELSA English Longitudinal Study of Ageing

F&Y / FYS Food and You Survey

FAO Food and Agriculture Organization

FOP Front-of-package (or Front-of-pack)

FSA Food Standards Agency

g Grams

GBM Generalized Boosted Models

GLM Generalized Linear Model

GLM-IPSW Generalized Linear Model-based Inverse Probability of Sampling Weighting

HSR Health Star Rating / Healthy Start Rating

IOSW Inverse Odds of Sampling Weighting

IPSW Inverse Probability of Sampling Weights

ITS Interrupted Time Series

kcal Kilocalories

KH Keyhole

MICE Multiple Imputation by Chained Equations

MTL Multiple Traffic Light

NDNS National Diet and Nutrition Survey

NEDV Non-Equivalent Dependent Variable

NHS National Health Service

NS Nutri-Score

OLS Ordinary Least Squares

OR Odds Ratio

PATE Population Average Treatment Effect

PSA Probabilistic Sensitivity Analysis

QALY Quality-Adjusted Life Year

RCT Randomized Controlled Trial

REM Ready-to-Eat Meals

RI Reference Intake

SATE Sample Average Treatment Effect

SD Standard Deviation

SE Standard Error

SUTVA Stable Unit Treatment Value Assumption

TMLE Targeted Maximum Likelihood Estimation

UI Uncertainty Interval

UK United Kingdom

UoM University of Manchester

UPF Ultra-processed foods

VIF Variance Inflation Factor

WL Warning Labels

Abstract

The escalating global prevalence of diet-related non-communicable diseases has spurred the broad implementation of Front-of-package nutritional labelling systems. Nevertheless, a pronounced disparity endures between the demonstrated efficacy of these informational interventions in controlled experimental contexts and their real-world effects across diverse national populations. This dissertation addresses this external validity shortfall by probing the dynamics of label prominence and cognitive limitations, while devising a rigorous causal inference methodology to extrapolate experimental results to the United Kingdom population. The inquiry comprises three interconnected empirical investigations. The initial study employs a quasi-experimental non-equivalent dependent variables approach, leveraging data from the United Kingdom Food and You Survey to disentangle the effects of label legibility from nutritional attributes. The second investigation, a randomized controlled trial, scrutinizes the interplay between label complexity and individuals' working memory capacity, assessed through an n-back task. To surmount the generalizability challenge, the culminating analysis introduces a novel transportability paradigm. This entails utilizing generative machine learning techniques to impute latent cognitive variables within the National Diet and Nutrition Survey, coupled with doubly robust Bayesian additive regression trees for estimation, thereby extrapolating trial-based causal effects to the population scale. These transported estimates feed into a dynamic ten-year microsimulation model to project enduring health and economic ramifications. The results contest the neoclassical precept that augmented nutritional disclosure invariably enhances dietary selections. Evidence from the first study posits label readability as a primary perceptual filter, exerting product-specific deterrent influences prone to temporal habituation. The second study unveils a cognitive heterogeneity dynamic: granular labels benefit those with superior working memory, yet overwhelm typical consumers' attentional limits. The transportability framework in the third study quantifies this selection bias

analytically, exposing a counterproductive rebound wherein intricate labels precipitate heightened caloric consumption population-wide, attributable to cognitive saturation and moral disengagement. Microsimulation forecasts suggest that isolated deployment of detailed labels incurs net adverse health consequences and substantial burdens on the National Health Service. In contrast, an integrated policy of rudimentary coarse-grained labels paired with compulsory industry reformulation proves optimal, averting hundreds of thousands of obesity incidents and generating considerable societal economic benefits. Methodologically, this work underscores the critical role of generative imputation and doubly robust methods in mitigating biases from unobserved effect modifiers in causal transport. Substantively, it affirms that equitable public health gains necessitate aligning label intricacy with prevailing cognitive capacities, emphasizing the imperative to couple streamlined consumer-facing signals with systemic supply-side interventions.

Lay abstract

Modern nutritional labels are designed to help consumers make healthier choices. However, most evidence comes from small trials that don't represent the whole country. This work uses advanced statistical methods to project how these labels would actually perform in the real world across the UK.

Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

i The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made *only* in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420`), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see `http://www.library.manchester.ac.uk/about/regulations/`) and in The University's policy on Presentation of Theses.

Acknowledgements

This thesis would not have been possible without the extensive help, guidance, and support of my supervisors, colleagues, friends, and family.

First and foremost, I would like to thank my supervisors, Yan Wang and Nick Shryane, for their unwavering support and guidance. I am profoundly grateful to them for introducing me to the field of quantitative evaluation of policies, interventions, and experiments. I also thank the acting editors, whose thorough reviews and numerous comments significantly improved the quality of the manuscripts.

I extend my sincere thanks to all the professors in the Department of Social Statistics at the University of Manchester for their continuous support throughout this program. I also express my gratitude to the participants in the RAGS Social Statistics seminars at the University of Manchester for their constructive comments and suggestions.

Part of this research was conducted while I was an Enrichment Student at the Alan Turing Institute. I thank the Institute for its welcoming environment and the highly inspiring discussions, particularly those within the Turing Research and Innovation Cluster in Digital Twins.

I owe special thanks to Diego Galaz and Tamara Ortega, whose academic contributions and suggestions have been invaluable in achieving the consistency required for this study to be a true contribution to society. I also wish to thank Javiera Ponce for her excellent research assistance.

On a personal level, I thank all the women who crossed my path and taught me how to wisely balance motherhood with my academic responsibilities. I am especially grateful to Chiara Ludolini and Fernanda Hernandez, my great friends who cared so deeply for my welfare during my PhD journey.

Finally, I am deeply grateful for the unconditional support of my family. I thank my mother and brothers, who have championed all my academic and life projects, as well as my sisters-in-law, nephews, niece, and mother-in-law. To my partner Franco, with whom I began and am now finally concluding this academic chapter, thank you for your boundless love, support, and trust. To my beloved daughter, who has accompanied me from my womb throughout the entirety of this dissertation—Franco and my baby, you have been and will always be my absolute inspiration.

The author

Constanza is a PhD researcher in Social Statistics at the Cathie Marsh Institute, University of Manchester (UoM) and Enrichment Student at the Alan Turing Institute. She holds a Master's degree in Social Research Methods and Statistics from the UoM. Constanza has led research initiatives encompassing the design and collection of data for extensive surveys, as well as the analysis of quantitative data. As head of the Department of Agricultural Studies at the National Institute of Statistics of Chile, Constanza conducted research on the food industry. These research projects have been highly valuable to the Central Bank of Chile and The Food and Agriculture Organization (FAO) of the United Nations in understanding macroeconomic growth dynamics. Constanza's doctoral research employs advanced machine learning techniques to examine the causal impact of front-of-package food labelling on consumer behavior, drawing on both randomized trials and observational data.

Artificial intelligence declaration

Generative Artificial Intelligence (AI) disclosure: I used Gemini 3.1 Pro to assist in idea generation, image creation, and for feedback on grammar and content.

Content notification

See dedicated policy at `https://documents.manchester.ac.uk/display.aspx?DocID=74816` for whether a content or trigger warning is required. Examples of such warnings are given in that policy.

Chapter 1

Introduction

1.1 The policy problem: dietary nudges, cognitive constraints, and public health impact

The global rise in diet-related diseases, such as obesity and cardiovascular conditions, poses one of the most significant and persistent public health challenges of the 21st century [1, 2, 3]. Worldwide, dietary risks were responsible for approximately 11 million deaths, with this trend particularly pronounced among lower-socioeconomic groups, thereby exacerbating health inequalities [4]. To mitigate these risks, front-of-package (FOP) nutritional labelling has emerged as a key public health strategy aimed at reducing the consumption of energy-dense, ultra-processed foods whose nutritional profile significantly impacts overall diet quality [1, 3]. Unlike back-of-package nutrition facts tables, FOP labels attempt to enhance consumer understanding of packaged foods through simplified symbols that quickly flag unhealthy products [5, 6]. Consequently, rigorously investigating the causal effects of this policy is a central component in advancing the management of diet-related diseases [7, 8, 9].

The fundamental principle of FOP labelling lies in disseminating accessible nutritional information to foster informed decision-making, thereby enabling consumers to better align their selections with long-term health objectives. From a classical economic perspective, FOP labelling represents a warranted policy intervention to inform individuals of health costs that are frequently not fully internalized at the point of purchase [9], given consumers' propensity for preference biases and underestimation of behavioral health consequences

[10, 11]. Labels constitute a nudge via disclosure-oriented policies [12], designed to rectify information asymmetries [11]. In doing so, this approach introduces transparency in food products composition [13] and respects consumer freedom of choice [14] without imposing stringent regulations.

Although mandatory and voluntary labelling policies have been extensively implemented worldwide, critical gaps remain regarding whether accessible information provision alone is necessary and sufficient to fundamentally change both immediate purchasing behaviors and long-term consumption patterns [7, 15, 8, 16]. Systematic reviews demonstrate that while randomized controlled trials consistently show that labels improve nutritional understanding in experimental settings, observational data from real-world implementations often yield attenuated, ambiguous, or null results [8]. This discrepancy suggests that individuals do not make suboptimal choices solely due to a lack of key information. Cognitive biases, default choices, and pre-existing food literacy—a critical foundational component that dictates whether informational nudges can successfully translate into actual consumption changes—actively shape consumer decision-making [17, 18, 19].

To fully grasp the mechanism and limitations of these interventions, it is crucial to clarify the conceptual distinction between consumer *behavior* and dietary *consumption* within the context of this thesis. Behavior encompasses the immediate, cognitive decision-making processes—such as visual attention, evaluating trade-offs, and the point-of-purchase intention—that occur in the shopping environment [7]. Consumption, conversely, refers to the subsequent, cumulative act of ingestion over time, which ultimately determines an individual's nutritional intake and long-term health outcomes [8]. FOP labels are primarily designed to intervene at the behavioral level by simplifying the choice architecture [12]. However, altering purchasing behavior does not automatically guarantee improved consumption, as compensatory behaviors (e.g., eating larger portions of a healthy labelled food) may occur [20]. Therefore, evaluating the true effectiveness of FOP policies requires understanding how these immediate behavioral nudges translate into sustained consumption changes.

FOP labels attempt to influence consumer behavior by updating individuals' underlying beliefs about a product's healthfulness. However, it is crucial to recognize that accurately interpreting a product's true nutritional quality is an ability largely restricted to individuals with a high baseline level of nutritional literacy [21]. Because this specific literacy is not a ubiquitous trait across the general population, the anticipated policy effects are inherently ambiguous. They hinge on consumers' prior beliefs about their shopping baskets, which may or may not be revised upon label exposure [13, 16]. Thus, even with full information, FOP labelling can sway choices in multiple directions, given its intrinsic connection to heuristic learning processes guided by personal beliefs [13, 16]. In other words, if an individual believes that Coca-Cola is healthier than water, exposure to a label may not necessarily correct this belief or lead to a change in purchase intentions.

From a psychological standpoint, the manner in which labeled information is presented plays a crucial role in prompting and updating belief mechanisms [11, 19]. Humans face an inherent trade-off between the time and effort required to locate and interpret information, owing to bounded attention and cognitive resources [12, 22, 23]. Consequently, the simplicity of the conveyed information emerges as a pivotal determinant [10]. We posit that for a label to effectively modify beliefs and thereby influence choices, it must prioritize such simplicity [10]. However, the literature has yet to converge on a consensus regarding the FOP labeling format that consumers deem most readable and comprehensible. The effects of discrete elements, such as colors and letters, have been extensively investigated, but key uncertainties persist concerning how the informational granularity of these formats—namely, the specific quantity of informational units provided—facilitates consumers' attention and interpretation [24, 15, 25].

The primary aim of this thesis is twofold: first, to evaluate how the cognitive ease of interpreting FOP labels influences consumer purchasing behavior, and second, to model how these specific behavioral shifts ultimately alter population-level consumption patterns through a formal causal inference framework. By investigating the mechanisms of label salience and the cognitive barriers to comprehension, this research seeks to

provide a nuanced understanding of how heterogeneity in consumer bandwidth shapes dietary decisions. Furthermore, the thesis aims to bridge the persistent gap between experimental evidence and real-world impact by developing and applying an advanced statistical framework to generalize trial findings to national target populations.

1.2 The UK Multiple Traffic Light policy: context and mechanisms

To ground this research within a specific structural and societal framework, this thesis centers on the United Kingdom's Multiple Traffic Light (MTL) policy. Introduced formally by the Department of Health and the Food Standards Agency (FSA) in 2013, the MTL represents a hybrid FOP labelling approach designed to standardize nutritional information across the UK food retail sector [26].

The MTL label combines reductive numerical information with interpretative, evaluative color-coding. Structurally, the label displays the absolute amounts of energy (calories), fat, saturated fat, sugars, and salt per portion or per 100 grams of the product. Crucially, the four specific nutrients (excluding overall energy) are color-coded—red (high), amber (medium), or green (low)—based on established nutritional thresholds [26].

The intended societal mechanism of the MTL relies on substantially reducing the cognitive friction associated with interpreting traditional back-of-package nutritional tables. By providing salient, at-a-glance visual cues, the policy operates through a dual mechanism to generate public health impact: altering consumer demand and stimulating supply-side product reformulation. On the demand side, the policy aims to disrupt habitual purchasing behaviors and nudge consumers toward healthier alternatives within the same product category. For instance, a shopper might quickly choose a ready-to-eat meal with predominantly green and amber labels over one displaying multiple red labels. On the supply side, as consumer preferences predictably shift toward these better-rated products,

the MTL inadvertently incentivizes food manufacturers to reformulate their offerings—such as by reducing salt, sugar, or saturated fat—to achieve a more favorable color profile, thereby improving the food environment systemically. Although the implementation of the MTL by food manufacturers and retailers remains officially voluntary, it has achieved widespread adoption, covering a significant majority of pre-packaged foods in UK supermarkets [9].

Furthermore, this policy design is intrinsically linked to addressing social inequalities in health [19]. Diet-related diseases and obesity disproportionately burden lower-socioeconomic groups [4]. Traditional, highly detailed nutritional tables often inadvertently widen these health disparities, as their interpretation relies heavily on advanced health literacy and numeracy skills that correlate with socioeconomic advantage. By utilizing a universally recognized heuristic—the traffic light color system—the MTL policy attempts to democratize nutritional information. The intent is to bypass complex cognitive requirements, making healthier food choices accessible and actionable for all consumers regardless of their educational or socioeconomic background, thereby aiming to narrow persistent health inequalities [26].

However, subsequent government consultations and empirical evaluations have highlighted persistent challenges in its practical application. While the MTL effectively flags unhealthy products via red warnings, its granular, multi-nutrient design still requires consumers to process several informational chunks simultaneously [15]. A shopper must often weigh complex trade-offs, such as comparing a product with a red sugar but green fat label against one with amber in both categories. This structural granularity means that the policy's effectiveness remains heavily contingent on the consumer's cognitive bandwidth and health literacy at the exact moment of decision-making [27].

Consequently, the UK MTL provides an ideal policy setting to explore the central aims of this thesis: evaluating the intersection of behavioral nudges, cognitive constraints, and the statistical challenge of estimating how these immediate point-of-purchase choices ultimately scale to population-level dietary consumption.

1.3 Methodological contribution: a tripartite causal inference framework

Evaluating the effectiveness of nutritional policies—specifically their capacity to meaningfully improve long-term, population-level consumption patterns—relies heavily on the quality of causal evidence. However, nutritional research frequently encounters severe methodological barriers when relying exclusively on either observational data or experimental trials. The availability of large-scale observational dietary data offers a glimpse into real-world behavior, but drawing causal inferences from these datasets is inherently complicated by unmeasured confounding. Dietary choices are inextricably linked to complex lifestyle patterns, socioeconomic status, and overall health consciousness. Because the true causal effect sizes of specific dietary components are often relatively small, they can be easily eclipsed by residual confounding and the substantial measurement errors typical of self-reported dietary assessment tools [28, 29]. Additionally, observational data often struggles with compound treatments; because the human diet operates within an isocaloric boundary, failing to explicitly define dietary compensations results in an ambiguous intervention, compromising the assumption of treatment consistency [30].

Alternatively, while a trial provides strong internal validity by eliminating baseline confounding through randomization, conducting long-term trials for sustained dietary behaviors is logistically difficult, expensive, and subject to declining adherence [29]. More critically, trial findings are routinely limited by a lack of external validity. Trial participants, who are typically self-selected volunteers, often possess higher baseline health literacy, motivation, and cognitive capacities than the broader public [31]. Consequently, the average treatment effect estimated within a trial may fail to accurately reflect the population average treatment effect required to inform national policymaking.

An ideal evaluation of a public health policy of this magnitude demands a comprehensive, tripartite approach: (1) observational analyses to capture real-world dietary patterns

and behavioral constraints, (2) experimental trials to isolate causal mechanisms with rigorous internal validity, and (3) generalization methods to securely project these experimental effects back onto the broader target population. Figure Fig. 1.1 illustrates this interconnected schema.

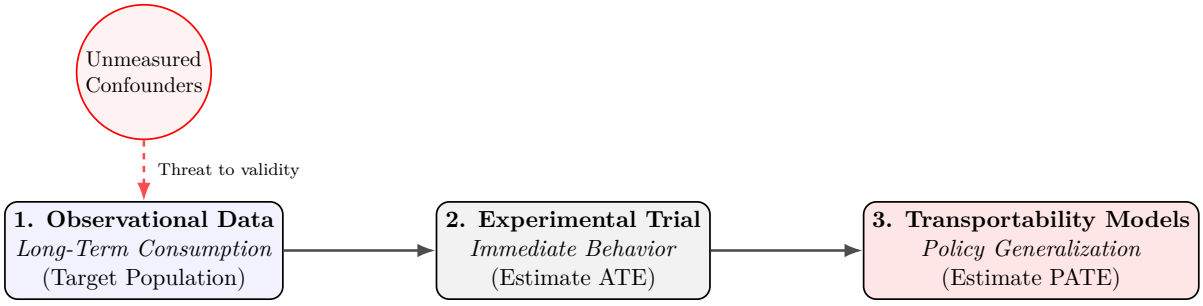

Fig. 1.1. Analytical trajectory of the tripartite causal inference framework. The methodology transitions from diagnosing the policy problem via observational *consumption* data (which is threatened by unmeasured confounders), to estimating local causal effects (ATE) on immediate purchasing *behavior* within internally valid experimental trials. Utilizing generalizability models bridges these two streams and compute the population average treatment effect (PATE).

By adopting this ambitious methodological synthesis, this thesis pioneers a highly rigorous causal inference framework. Moving beyond standard unadjusted estimates that often characterize public health research, it highlights a deep methodological originality: the capacity to not only assess the immediate experimental efficacy of dietary interventions but to statistically project their real-world, population-level impact. Crucially, the value of this tripartite methodology extends far beyond nutritional epidemiology. By formalizing a pipeline from observational diagnosis to experimental identification and, ultimately, causal generalization, this thesis provides a highly adaptable and replicable template. Other researchers can readily translate this framework to rigorously evaluate complex interventions across diverse fields—ranging from behavioral economics to broader social and environmental policies—ensuring that future evidence synthesis is both internally robust and externally valid.

1.3.1 Thesis outline

Situated within this interdisciplinary landscape, this thesis investigates FOP labelling effectiveness by addressing both the empirical constraints of individual cognition and the methodological challenges of integrating data to estimate population-level impacts. The remainder of this thesis is structured across four subsequent chapters:

- **Chapter 2: Theoretical and methodological background.** This chapter outlines the theoretical and methodological background of the thesis. It details the potential outcomes framework, the structural assumptions required for causal transportability, and the use of doubly robust machine learning estimators to integrate data sources while minimizing model misspecification.
- **Chapter 3: Food label readability and consumption frequency: isolating content-specific effects via a non-equivalent dependent variable design.** Directly linking to the aim of investigating mechanisms of label salience, this chapter explores the salience-to-understanding pathway using observational data. Acknowledging the challenges of unmeasured confounding in nutritional epidemiology, it applies a quasi-experimental non-equivalent dependent variable design to repeated cross-sectional UK survey data. It investigates whether subjective perceptions of label print size readability block the effectiveness of nutritional warnings on ultra-processed ready-to-eat meals, isolating content-specific effects from general label presence. By highlighting that perceptual clarity is a critical bottleneck for less-engaged consumers, this chapter connects the physical limitations of label design to the behavioral realities of consumption.
- **Chapter 4: Food label granularity and working memory: effects on food choice in a randomized controlled trial.** Building on the need to understand cognitive barriers to comprehension, this chapter utilizes an experimental trial to generate internally valid causal estimates regarding label design. Grounded in

cognitive-load theory, it tests how label granularity interacts with individual cognitive constraints. The analysis examines whether the effectiveness of detailed versus coarse nutritional labels on calorie selection is actively modified by a consumer's working memory capacity. While consumption decisions are ultimately driven by a multitude of cultural, environmental, and economic factors, this chapter specifically isolates the role of cognitive constraints and informational granularity to test the hypothesis that label effectiveness increases with available cognitive bandwidth.

- **Chapter 5: On the selection of covariates for transportability: a doubly robust BART approach for generalizing a food labeling trial and modeling long-term public health impact.** Finally, addressing the aim of bridging the gap between experimental evidence and real-world impact, this chapter synthesizes the preceding analyses by applying transportability methods to a non-nested design. It integrates the experimental trial data from Chapter 4 with observational data from the UK National Diet and Nutrition Survey. Using doubly robust Bayesian additive regression trees (BART), this chapter transports the local trial behavioral effects to estimate the national population average treatment effect on consumption. Furthermore, it incorporates these estimates into a microsimulation to model long-term policy cost-effectiveness and utilizes sensitivity analyses to explicitly quantify the robustness of the estimates against potential omitted covariates.

Each of the empirical chapters (Chapters 3, 4, and 5) is structured as a self-contained research paper, complete with specific introductions, methodologies, and conclusions. While the chapters can be read independently, the prescribed sequence reflects the methodological progression of the thesis: diagnosing behavioral constraints via observational data, identifying causal mechanisms experimentally, and utilizing generalizability frameworks to forecast national public health impact.

Chapter 2

Literature review: a causal inference framework for policy generalization

2.1 Introduction

Evidence-based public health policymaking relies on a pivotal question: can interventions proven effective in controlled experimental settings yield equivalent benefits when scaled to a heterogeneous national population? Although randomized controlled trials (RCTs) are universally considered the gold standard for establishing local causal effects, they are frequently undermined by a fundamental trade-off between internal and external validity. The stringent inclusion criteria, artificial environments, and self-selecting volunteer pools that grant trials their high internal validity simultaneously compromise their external validity, leaving policymakers uncertain about real-world efficacy [32, 31].

This chapter elucidates the theoretical and methodological literature essential for bridging this external validity gap. By establishing a tripartite causal inference framework, it methodically delineates and examines the foundational concepts underpinning causal effect transportability—encompassing potential outcomes, treatment effect heterogeneity, unmeasured confounding, and the statistical complexities inherent to non-nested designs. Through this synthesis, the chapter constructs a rigorous, reproducible framework for assessing immediate behavioral responses and projecting the long-term, real-world consumption effects of dietary interventions.

2.2 Nutrition and public policy

According to the theoretical framework advanced by Grunert and Wills, an effective front-of-package (FOP) label must be noticed, appreciated, and understood [21]. Empirical evidence demonstrates that labels bolster consumers' capacity to identify healthier products when the conveyed information is objectively understood, thereby enabling precise distinctions based on complex nutritional profiles [6, 33, 34, 35, 36, 15]. Furthermore, such labels enhance product selection abilities when individuals subjectively perceive comprehension of the nutritional details, highlighting the critical role of the salience-to-understanding pathway [6, 33, 34, 37]. However, this understanding is not universal; it is heavily modulated by demographic traits, personal interest, baseline health literacy, and the specific granular design of the label [21, 38].

Nutritional labels aim to fundamentally alter the choice architecture of the food environment. They act as informational nudges—serving to inform rather than comprehensively educate—that simplify complex nutritional data into salient, at-a-glance evaluations to empower healthier point-of-purchase consumer behavior [8, 7]. Yet, evaluating the genuine efficacy of FOP labeling policies is notoriously difficult at the population level. Systematic reviews demonstrate a persistent paradox: while experimental trials consistently show that labels improve nutritional understanding and reduce caloric selection in controlled settings, observational data from real-world implementations often yield attenuated, ambiguous, or null results [8]. This discrepancy is deeply rooted in behavioral heterogeneity and cognitive constraints, as the cognitive bandwidth required to process detailed labels is unevenly distributed across the population [39].

This discrepancy also highlights the fundamental methodological obstacles in nutritional research. Conducting long-term trials to track sustained dietary consumption patterns is often logistically impractical, excessively costly, and hindered by declining participant adherence over time [29]. Consequently, researchers must frequently rely on observational data to evaluate long-term impacts. However, drawing robust causal conclusions from

observational nutritional data is severely complicated by unmeasured confounding. Dietary choices are inextricably linked to complex lifestyle patterns, socioeconomic status, and overall health consciousness [28]. Because the true causal effect sizes of specific dietary interventions are often relatively small, they can be easily eclipsed by residual confounding and the substantial measurement errors inherent in self-reported dietary assessment tools [28, 29].

Collectively, these challenges emphasize the absolute necessity of moving beyond naive observational correlations. Estimating the true impact of nutritional policies requires a formal causal framework capable of integrating distinct data sources. This involves leveraging the rigorous internal validity of experimental trials concerning immediate purchasing behavior, while systematically addressing their limited external validity by mapping these effects onto the complex covariate structure of real-world observational consumption data.

2.3 Causal inference framework for food labeling intervention

To rigorously evaluate policy impact across the translational gap from behavior to consumption, methodological research adopts a formal causal inference modelling approach rooted in the Neyman-Rubin potential outcomes framework [40]. The fundamental problem of causal inference is that we can only observe one reality for any given individual. Let A denote a binary treatment assignment, where $A = 1$ represents exposure to a detailed FOP label and $A = 0$ represents a control condition (e.g., a coarse label or no label). For each subject i, there exist two potential outcomes, $Y_i(1)$ and $Y_i(0)$, representing the outcome (e.g., calories selected) under treatment and control, respectively.

The individual causal effect is defined as the difference $Y_i(1) - Y_i(0)$. Because it is empirically impossible to observe both potential outcomes simultaneously for the same

individual, it is necessary to estimate average causal effects over a specified population. This framework critically relies on the stable unit treatment value assumption (SUTVA), which posits two conditions: first, there is no interference between subjects (one person's label exposure does not affect another shopper's outcome), and second, the treatment is consistently defined without hidden variations or ambiguous dietary compensations [40].

Within the controlled environment of a trial, the treatment indicator A is randomly assigned, rendering it independent of both observed and unobserved baseline covariates X (i.e., $X \perp A$). This randomization process establishes strict exchangeability between the treatment and control arms, permitting the unbiased estimation of the causal effect through the direct comparison of their mean outcomes. Randomization thus securely establishes the internal validity of the trial. However, despite this robust causal identification, the resulting estimand remains fundamentally local to the trial sample, underscoring the urgent need to distinguish among various average treatment effects when projecting broader public health impact [31].

2.4 Generalizing trial findings to target populations

A national food labeling policy aims to improve public health equitably across the entire population, not merely among the self-selected participants of experimental cohorts. Therefore, the population of substantive policy interest is the target population—the comprehensive, real-world collective of individuals eligible for the intervention [41].

To bridge the methodological divide between the trial sample and the target population, it is essential to mathematically differentiate between two interrelated concepts in the causal inference literature: generalizability and transportability [32]. Generalizability applies exclusively to nested designs, where the trial sample constitutes a direct, probability-based random subset of the target population registry, enabling a straightforward extrapolation of findings back to that source population. In contrast, transportability is pertinent to

non-nested designs, where the trial sample and target population are distinct and entirely non-overlapping. This scenario is overwhelmingly prevalent in public health research, where behavioral results from a targeted, volunteer-based experiment must be extrapolated to a separate national health survey representing the broader, long-term consumption patterns of the population [32, 31]. The structural differences between these horizontal designs are illustrated in Figure Fig. 2.1.

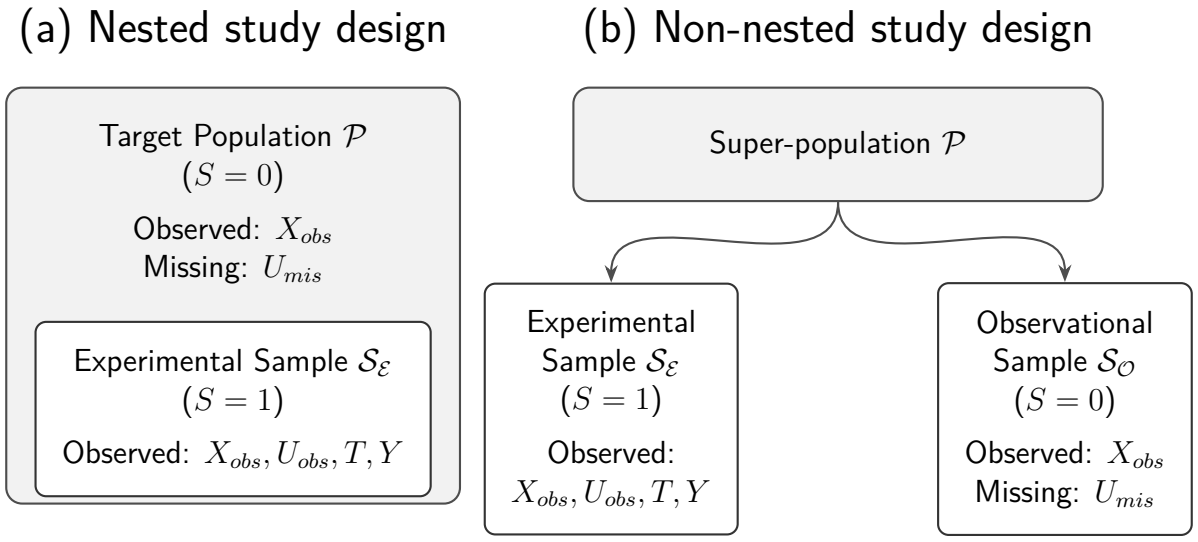

Fig. 2.1. Comparison of nested and non-nested sampling designs for transportability arranged horizontally. Schematic representation adapting the causal framework of Colnet et al. Panel (a) illustrates a nested study design, where the experimental sample is embedded directly within the target population registry. Panel (b) illustrates a non-nested study design, where an experimental sample and an observational target sample are drawn independently from a super-population $\mathcal{P}$. In the experimental sample ($S = 1$), universally observed covariates (X_{obs}), partially observed effect modifiers (U_{obs}), assigned treatment (T), and outcomes (Y) are fully observed. Conversely, in the observational target sample ($S = 0$), only universally observed covariates (X_{obs}) are recorded, while critical effect modifiers (U_{mis}) are systematically unobserved.

Both methodologies become statistically intricate in scenarios involving heterogeneous treatment effects—namely, when the label's influence is actively moderated by particular covariates. For instance, if a detailed food label exerts a strong positive influence on individuals with high cognitive bandwidth and working memory, but has minimal or counterproductive impacts on those with low cognitive capacity, then working memory functions as a critical effect modifier [42]. Should the trial sample disproportionately

over-represent highly literate, cognitively advantaged individuals relative to the target population, a simplistic analysis would substantially overestimate the policy's true systemic benefits. Accordingly, transporting trial results to accurately forecast public health impact demands reweighting or standardizing the experimental data to precisely align with the target population's nuanced covariate distribution [43]. Table Table 2.1 summarizes the methodological differences between these two frameworks across all introduced core concepts.

2.5 Population average treatment effects: identification and assumptions

To formalize transportability, it is essential to mathematically differentiate the causal effect observed within the experimental sample from the projected effect in the broader target population. Define S as a binary indicator of trial participation, where $S = 1$ denotes inclusion in the trial and $S = 0$ denotes exclusion therefrom. The conventional estimand yielded by a randomized trial is the sample average treatment effect (SATE), defined as the average causal effect conditional on trial participation:

$$\tau_{SATE} = \mathbb{E}[Y(1) - Y(0) \mid S = 1] \tag{2.1}$$

While possessing robust internal validity, the SATE is of severely limited utility to policymakers if $S = 1$ represents a highly skewed or unrepresentative demographic. The policy-relevant estimand is the population average treatment effect (PATE). PATE represents the average causal effect expected if the entire target population were systematically assigned to the treatment versus the control:

$$\tau_{PATE} = \mathbb{E}[Y(1) - Y(0) \mid S = 0] \tag{2.2}$$

Table 2.1. Methodological comparison of Generalizability and Transportability designs. Theoretical framework adapted from Colnet et al. (2022)

Design characteristic	**Generalizability (Nested design)**	**Transportability (Non-nested design)**
Target population	A defined, finite observational registry or cohort serving directly as the target population.	An infinite or large underlying super-population from which distinct samples are drawn.
Experimental sample	Individuals are selected directly from the established target registry ($S = 1$).	Individuals are drawn independently from the super-population to form the trial ($S = 1$).
Observational sample	The remaining, non-randomized individuals naturally present within the identical registry ($S = 0$).	A separate, independent survey sample drawn from the broader super-population ($S = 0$).
Target estimand	Population average treatment effect (PATE) across the defined finite registry.	Population average treatment effect (PATE) across the underlying super-population.
Ignorability (Exchangeability)	Requires measuring all confounders that predict trial inclusion *within* the established registry.	Requires measuring all effect modifiers (X, U) that differ between the independent trial and survey datasets.
Positivity (Overlap)	All individuals in the registry have a known, non-zero probability of trial selection.	All target population profiles must have a non-zero probability of representation in the trial.
Data architecture	A single, unified dataset containing both trial participants and non-participants.	Two distinct, unlinked datasets that must be mathematically harmonized.
Sampling probability	The absolute probability of trial participation, $P(S = 1 \mid X)$, is directly identifiable.	Absolute probability is non-identifiable; estimation relies strictly on the relative density ratio or inverse odds.

Whenever key effect modifiers are distributed differently between the trial ($S = 1$) and the target population ($S = 0$), the SATE will not equal the PATE ($\tau_{SATE} \neq \tau_{PATE}$). Estimating the true PATE requires identifying a sufficient set of covariates X that comprehensively account for both the selection mechanism into the trial and the modification of the treatment effect [44].

To rigorously identify the PATE from trial data and a target population survey, a series of strict causal assumptions must hold [32, 31]. Beyond the standard assumptions of randomized trials, transportability demands two primary structural assumptions:

1. **Conditional exchangeability over selection (ignorability):** This assumption states that, conditional on a sufficient set of observed pre-treatment covariates X, participation in the trial S is independent of the potential outcomes:

$$Y(a) \perp S \mid X \quad \text{for } a \in \{0, 1\} \tag{2.3}$$

   In practice, researchers must measure and condition on all covariates that affect both the probability of trial participation and treatment effect heterogeneity. Failure to measure a key effect modifier (e.g., unmeasured cognitive capacity) violates this assumption, yielding directionally biased PATE estimates.

2. **Positivity of trial participation (overlap):** This assumes that for every combination of covariates observed in the target population, there is a non-zero, positive probability of those individuals being included in the trial sample:

$$P(S = 1 \mid X = x) > 0 \quad \text{for all } x \text{ where } P(S = 0 \mid X = x) > 0 \tag{2.4}$$

   If the target population includes a specific demographic group that was completely excluded from the experimental trial, the causal effect cannot be securely transported to that sub-group without relying on highly speculative and untestable mathematical extrapolation.

2.6 Estimation strategies: single and doubly robust methods

2.6.1 Parametric approaches: weighting and outcome modelling

Traditionally, the generalization of randomized trial results to target populations has employed two primary parametric methods: weighting and outcome modeling. The weighting approach, typically implemented via inverse probability of sampling weights (IPSW), explicitly models the trial selection mechanism. A parametric logistic regression is usually fitted to the concatenated dataset to estimate $P(S = 1 \mid X)$, the probability of trial participation given baseline covariates. Trial participants are then reweighted by the inverse odds of their selection, thereby increasing the influence of those resembling the target population while significantly reducing the weight of overrepresented subgroups [31, 45]. The PATE is subsequently estimated as the weighted mean difference in outcomes between treated and control groups in the trial.

In contrast, outcome modeling—often referred to as G-computation or standardization—disregards the trial selection mechanism entirely. It entails fitting a parametric regression model—such as linear regression for continuous outcomes or logistic regression for binary outcomes—to the trial data in order to predict the outcome Y as a function of the treatment A and covariates X. This model is then applied directly to the target population survey data to estimate counterfactual outcomes $Y(1)$ and $Y(0)$ for each individual therein. The PATE is thereafter calculated as the mean of these predicted treatment effect differences across the entire target dataset.

Although theoretically justified, both of these parametric methods remain severely vulnerable to a common statistical pitfall: model misspecification. Should the analyst assume a simplistic linear or logistic functional form, while the true behavioral process exhibits deep nonlinearity or intricate multi-way interactions among covariates—such as the interaction

between cognitive capacity, socioeconomic status, and baseline diet—the resulting models will be misspecified, inevitably producing substantially biased PATE estimates.

2.6.2 Doubly robust estimators

To mitigate the pervasive risks of model misspecification, modern causal inference heavily utilizes doubly robust (DR) estimators (such as Augmented IPSW). DR estimators mathematically combine both the selection model (weighting) and the outcome model (standardization). Their defining and highly advantageous property is that if either the selection model or the outcome model is correctly specified, the resulting PATE estimate remains unbiased and mathematically consistent [46, 31].

Even though DR estimators can be implemented via conventional parametric regressions, their full analytical potential is only realized when seamlessly integrated with nonparametric generative machine learning methods. Employing algorithms such as Bayesian additive regression trees (BART) allows researchers to systematically surmount the constraining assumptions of parametric models [47, 48]. BART constitutes a versatile, Bayesian machine learning ensemble technique that aggregates decision trees, inherently accommodating complex interactions and non-linearities without mandating explicit functional form specifications [49]. Integrating BART into a doubly robust transportability framework delivers a highly flexible, data-adaptive machine learning approach that accepts the variance penalty inherent in algorithmic imputation in exchange for optimizing the prospects for correct model specification, proving statistically superior to relying on flawed demographic proxies [50, 51].

2.7 Sensitivity analysis for omitted covariates

Transportability and generalizability necessitate the complex integration of two distinct data sources: randomized experimental behavioral data alongside observational data, such

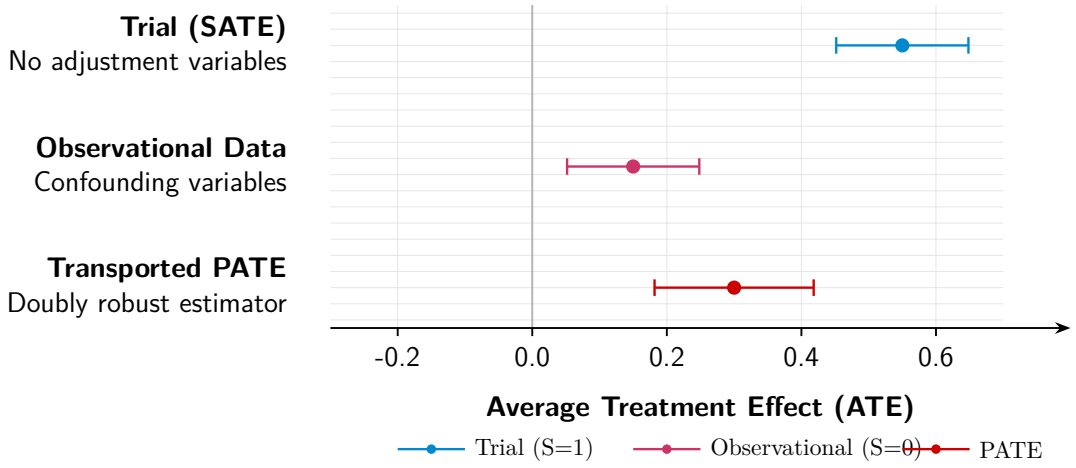

Fig. 2.2. Juxtaposition of different estimation results for educational purposes, adapting the visual framework of Colnet et al. (2022). The plot illustrates how naive randomized controlled trials (RCT) can systematically overestimate the Average Treatment Effect (ATE) when applied directly to a target population. Conversely, pure observational estimates derived from national surveys may be severely biased by unmeasured confounding despite adjusting for many variables. The transported Population Average Treatment Effect (PATE) effectively uses doubly robust machine learning to bridge these designs, adjusting the trial estimate based on the target population's true covariate distribution to generate a more accurate, real-world causal effect.

as representative consumption surveys or national health records. A core threat to the validity of transportability analyses arises from violations of the conditional exchangeability assumption due to unmeasured covariates. As previously outlined, successfully transporting the SATE from a trial to a target population strictly requires the observation of all relevant treatment effect modifiers in both the experimental and survey datasets. Yet, in intricate, multifaceted domains such as nutritional behavior and public health, it remains highly likely that key behavioral, socioeconomic, or cognitive traits—such as digital literacy, baseline nutritional anxiety, or culturally ingrained food preferences—go unmeasured in one or both sources [52, 53].

If an unobserved variable simultaneously influences an individual's probability of participating in the trial and the magnitude of the treatment effect they experience, the assumption of conditional mean exchangeability is broken. In such scenarios, the resulting PATE estimate will be directionally biased, potentially masking counterproductive behavioral outcomes like moral licensing or compensatory eating, and ultimately leading

to erroneous policy recommendations [31].

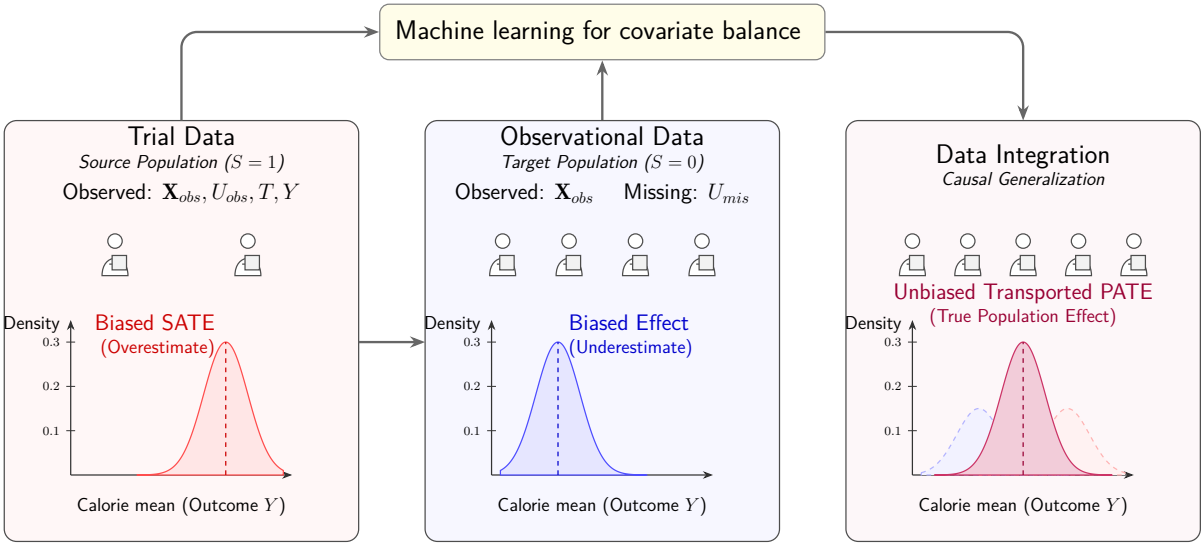

Fig. 2.3. Causal inference framework for generalizing trial findings, arranged horizontally. $S = 1$ denotes the source experimental sample, and $S = 0$ denotes the target national population. $\mathbf{X}_{obs}$ represents the vector of universally observed pre-treatment covariates. U_{obs} represents variables (like working memory scores) fully observed within the trial, while U_{mis} denotes the unmeasured cognitive capacity in the target survey. T indicates the assigned label treatment, Y denotes the caloric choice outcome, and $\hat{\tau}_{DR}$ represents the transported population average treatment effect. As demonstrated, the generative machine learning model operates by learning to correct the covariate distribution of the trial data from the observational data to detect covariate imbalance. By formally learning the true covariate distribution present in the general population, the machine learning algorithm adjusts the trial covariate distribution to generate unbiased, real-world estimates despite missing parameters.

To systematically address this inevitable uncertainty, modern causal inference relies on formal sensitivity analyses [52]. Rather than resting on the fragile, untestable assumption that ignorability holds perfectly, sensitivity analysis allows researchers to mathematically quantify exactly how robust their transported estimates are to potential unmeasured confounding. By systematically varying theoretical parameters that represent the strength of a hypothetical unobserved confounder's relationship with both trial selection and the outcome, it becomes possible to determine the specific threshold at which the policy-relevant conclusions would change direction or lose statistical significance [54, 53]. Incorporating these sensitivity bounds provides a transparent and rigorous mechanism to handle omitted covariates, ensuring that final public health estimates are evaluated with appropriate

methodological caution.

2.8 Software implementations for generalizability

The swift evolution of transportability methodologies has coincided with the emergence of open-source software designed to support both the theoretical identification and practical estimation stages of analysis. Table Table 2.2 presents a concise overview of these tools, adapted from Colnet's comprehensive methodological review [31]. Identification-oriented software primarily emphasizes the algorithmic evaluation of causal graphs and selection diagrams, whereas estimation software implements weighting, G-computation, and doubly robust estimators, accommodating both traditional parametric frameworks and advanced machine learning approaches.

Table 2.2. Summary of publicly available software for the generalization of trial findings (Adapted from Colnet, 2024).

Phase	**Methodological Task**	**Software / Packages**
Identification	Non-parametric identification of causal effects via do-calculus, selection diagrams, and causal graphs (DAGs).	`causaleffect`, `dosearch`, `dagitty`, `DoWhy`
Estimation	Implementation of Inverse Probability/Odds Weighting (IPSW/IOSW), G-computation, and Doubly Robust estimators using parametric models or Machine Learning.	`extRCT`, `geex`, `AIPW`, `SuperLearner`, `BART`

2.9 Summary

Translating experimental evidence into genuine public health impact requires far more than rigorous trials; it demands a formal, robust framework capable of extending local behavioral efficacy findings to model long-term, population-level consumption effects. The

integration of the Neyman-Rubin potential outcomes framework with transportability methodologies directly addresses key external validity gaps in food labeling research. By explicitly accounting for variations in cognitive bandwidth, health literacy, and dietary compensations outside controlled laboratory settings, the use of non-nested survey designs and doubly robust machine learning estimators enables the secure transport of trial estimates. This tripartite framework ultimately provides a powerful structure for interpreting experimental findings, rigorously addressing unmeasured confounding, and estimating population average treatment effects to design and inform truly equitable national nutritional policies.

Chapter 3

Food Label Readability and Consumption Frequency

Abstract

Objective: This study investigates the association between consumers' perceived readability of Multiple Traffic Light (MTL) label print size—a theoretical structural gatekeeper for visual salience—and self-reported food consumption frequency in the United Kingdom. We aimed to disentangle the effect of label readability from label content. Using non-equivalent dependent variables (NEDVs), we tested whether the association is specific to unhealthy convenience foods and absent for healthy or unlabeled foods, while also examining heterogeneity across consumer subgroups. Methods: Data from 8948 adults across four waves (2012–2018) of the UK Food and You Survey were analyzed. Cumulative link ordinal logistic regressions were employed to model the association between self-reported print size readability and the consumption frequency of four product types: pre-packaged sandwiches and pre-cooked meat (unhealthy, labeled targets), dairy (nutritionally advisable, labeled control), and fresh meat (unlabeled control). Models were adjusted for sociodemographic covariates, health behaviors, and survey wave fixed effects. Results: The findings reveal a content-specific and significant dynamic relationship exclusively for pre-packaged sandwiches. In 2012, a one-unit increase in readability was associated with a 9% decrease in the odds of frequent consumption ($OR = 0.91$), consistent with a warning effect. However, by 2018, this relationship reversed to a 4% increase ($OR = 1.04$), indicating that higher readability became associated with more frequent consumption. In contrast, a persistent null association was observed for pre-cooked meat, dairy, and fresh meat. Subgroup analyses for sandwiches indicated that the association with readability was strongest among less-engaged consumers. Conclusions: Empirical evidence challenges the utility of a standardized approach to food labelling. The results suggest that the effectiveness of label salience is contingent not just on the consumer but on the product's context and the content of its message, highlighting the need for adaptive rather than uniform policy standards.

3.1 Introduction

Diet-related health issues remain a critical challenge in the United Kingdom. Ultra-processed foods (UPFs) now account for over half of dietary energy intake [55, 56]. Among these, Ready-to-Eat Meals (REMs) represent a rapidly growing category. This sector has achieved 90% market penetration [57]. Unlike snacks or ingredients, REMs are consumed as complete meal replacements [58, 59]. Consequently, their nutritional profile significantly impacts overall diet quality. However, convenience often masks poor nutritional quality. REMs frequently contain high levels of saturated fat, sodium, and sugar [60, 61], increasing cardiovascular disease risk [62]. High consumption frequency and nutritional homogeneity make REMs an ideal case study. They are well-suited for evaluating front-of-pack (FOP) labelling interventions aimed at signaling health risks.

The UK government introduced the Multiple Traffic Light labelling policy in 2013 (Figure Fig. 3.1). Self-reported understanding is high [38]. Yet market data reveal only modest shifts in REM purchasing behavior [19]. A distinction between label presence and label readability offers a plausible explanation. Current UK Food Standards Agency guidelines mandate a minimum font size of just 1.2 mm [63]. A label may be visually salient due to color. Nevertheless, it fails as a communicative tool if the text falls below a consumer's perceptual threshold.

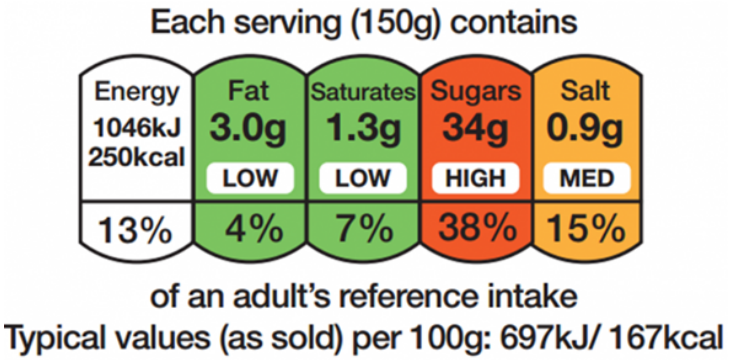

Fig. 3.1. MTL labelling visual representation. MTL displays the percentages of the recommended daily nutrient intake as the numerical assessment of the product's overall contribution for an average adult diet of 2000 calories.

Clarifying label effectiveness requires distinguishing four theoretical constructs. Salience refers to initial attentional capture, often driven by color or size [64]. Readability describes the perceptual ease of decoding text [65]. Objective understanding measures the accuracy of nutritional interpretation [66]. Finally, subjective understanding is the consumer's feeling of being informed [37]. Objective understanding is crucial for nutritional literacy. However, consumer choice models suggest that subjective understanding is often the more proximal driver of purchasing decisions [21]. Color generates salience, but readable text is necessary for cognitive decoding. If text falls below the perceptual threshold, the pathway to subjective understanding is blocked. This renders the warning ineffective regardless of visual prominence.

We propose a salience-to-understanding mechanism to synthesize this process (Figure Fig. 3.2). To influence behaviour, a label must traverse a pathway from visual detection to subjective understanding [6]. This feeling of being informed predicts choice. Theoretical frameworks suggest this pathway is not uniform. It is likely moderated by consumers' pre-existing health beliefs and motivations [67]. The Health Belief Model posits that highly engaged consumers are already sensitized to risk information. They may use labels merely to confirm their choices [68]. In contrast, perceptual clarity may be the critical bottleneck for less-engaged consumers. Without high readability, the warning signal may be ignored. Thus, we expect substantial heterogeneity across consumer subgroups.

Evaluating this mechanism poses a methodological challenge: distinguishing readability from content. A legible label does not intrinsically deter consumption. Rather, it intensifies the conveyed message. For unhealthy REMs, enhanced readability should strengthen the warning signal and theoretically curb consumption. In contrast, readability may reinforce a positive signal for healthier products [69]. This could boost consumption or leave it unaffected. Conventional observational analyses may confound these opposing effects. Moreover, standard quasi-experimental methods like Difference-in-Differences (DiD) prove impractical. The policy's simultaneous nationwide rollout lacks arbitrary exposure assignment rules [70].

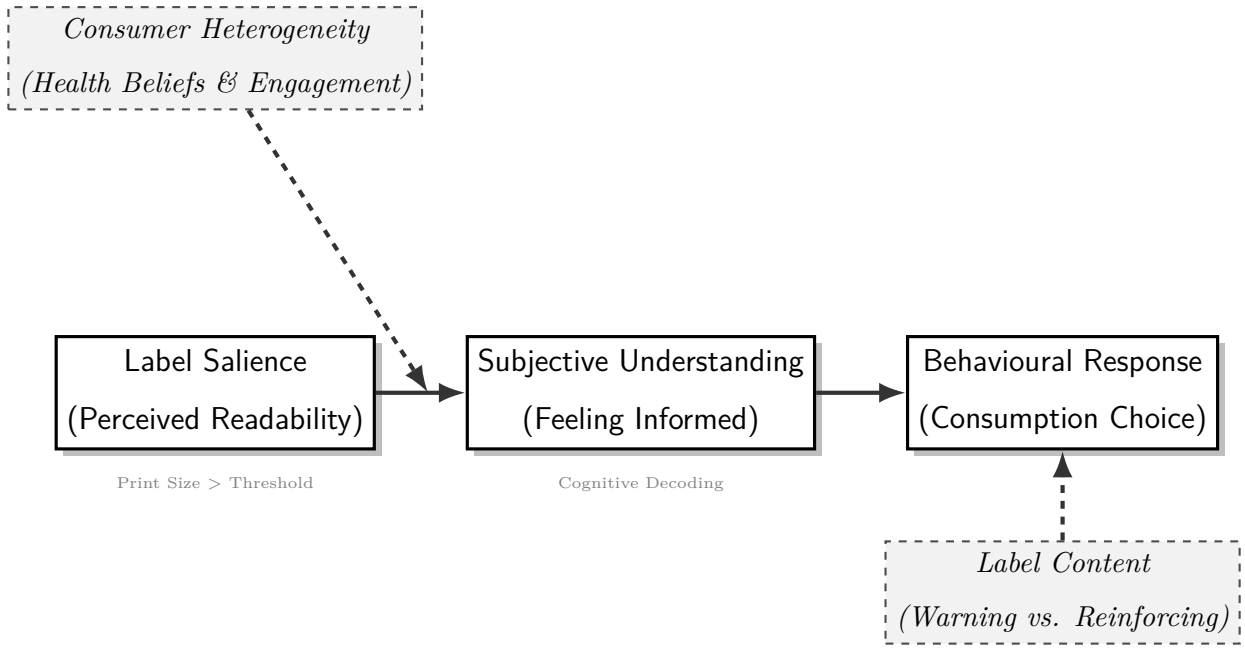

Fig. 3.2. The salience-to-understanding conceptual framework. The model illustrates the hypothesized pathway where label readability (salience) facilitates subjective understanding, which in turn influences consumption behavior. This pathway is moderated by consumer heterogeneity (health beliefs and engagement) and the specific content of the label (warning vs. reinforcing signals).

This study implements a Non-Equivalent Dependent Variable (NEDV) design to overcome these challenges [71]. We leverage repeated cross-sectional data from the UK Food and You Survey. We isolate the effects of label readability by contrasting target outcomes (unhealthy REMs) with control outcomes (dairy and fresh meat). These controls share exposure to common market trends. However, they remain theoretically insulated from the warning mechanism. Dairy typically features green/amber codes, while fresh meat carries no labels.

This research investigates the association between perceived readability of MTL print size and food consumption frequency. It specifically tests the hypothesized salience-to-understanding pathway through three hypotheses:

- H1: Perceived readability will have a significant, dynamic association with the consumption of unhealthy REMs (sandwiches/pre-cooked meat), reflecting a response

to the warning signal.

- H2: Perceived readability will show a null or reinforcing association for the healthy labeled control (dairy) and a null association for the unlabeled control (fresh meat). This confirms the effect is not a result of general consumer vigilance.
- H3: The association between readability and consumption will be strongest among less-engaged consumer subgroups, for whom label salience acts as a primary informational cue rather than a confirmation of pre-existing beliefs.

3.2 Materials and Methods

3.2.1 Study Design

This study employs a quasi-experimental Non-Equivalent Dependent Variable (NEDV) design (Figure Fig. 3.3) [71, 72, 73]. A primary challenge for evaluating this policy was that the MTL intervention was implemented across the entire United Kingdom simultaneously in 2013. Because the whole population was exposed at once, it was impossible to establish a conventional control group of unexposed individuals [74].

Consequently, standard quasi-experimental methods that rely on comparing treated and untreated populations, such as Difference-in-Differences (DiD), were not feasible [70]. Similarly, an Interrupted Time Series (ITS) design was rejected due to the low temporal resolution of the available biannual data, which precludes the robust modelling of trends required to detect a specific policy step change [75]. To overcome these limitations, we adopted the NEDV design, which constructs a valid counterfactual using outcome variables (comparable food products) rather than population groups (people) [72].

A NEDV is an outcome variable theoretically insulated from the intervention but responsive to potential confounders, such as local history or selection bias [76, 77]. Causal inference

is strengthened if the effect manifests in the target dependent variable but is absent in comparable NEDVs, thereby ruling out generic biases [78].

The target outcome is the consumption frequency of REM products (pre-packaged sandwiches and pre-cooked meat), which are theoretically responsive to red warning labels [61, 79]. We selected two NEDVs to control for market and behavioral confounders. Dairy products (labeled control) share similar retail environments and industry strategies (e.g., reformulation) but typically display reinforcing green/amber codes rather than warnings [13, 55]. Fresh meat (unlabeled control) captures general consumer trends but remains isolated from the specific MTL mechanism [80]. We acknowledge two validity threats: differential history (category-specific market changes) and mechanism diffusion (generalized caution toward all labels) [68, 81, 82]. Testing consistency across these varied controls mitigates these threats.

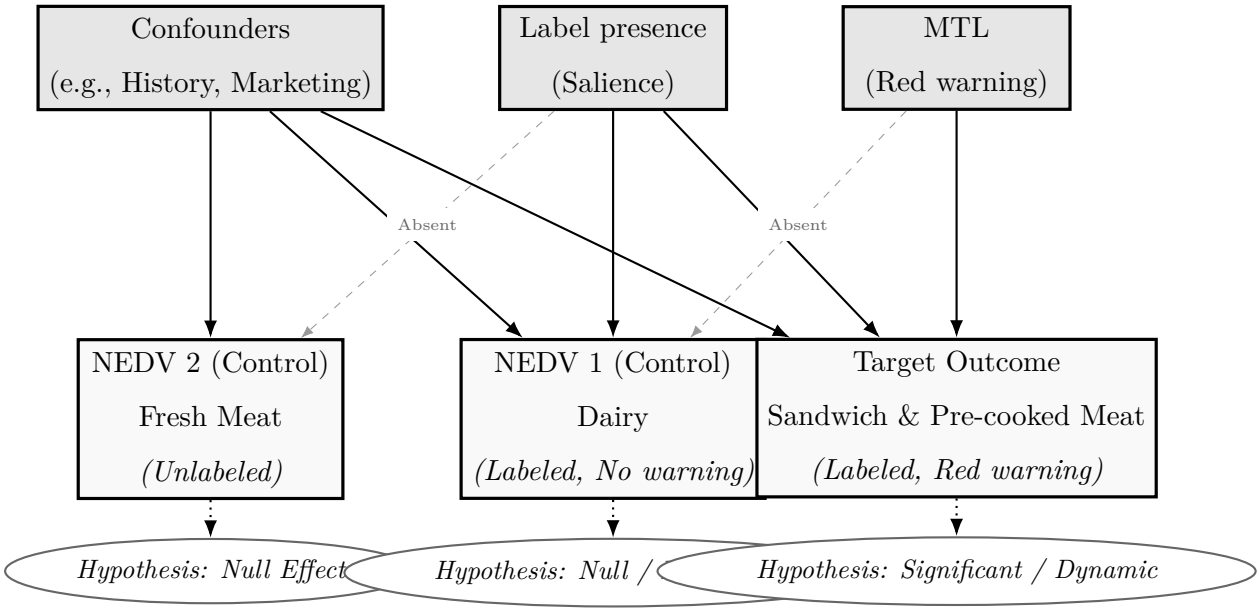

Fig. 3.3. Conceptual logic of the Non-Equivalent Dependent Variable (NEDV) design. The design isolates the effect of the MTL warning mechanism by comparing the target outcome against controls that share general environmental confounders but lack the specific intervention mechanism. Solid lines represent exposure presence; dashed lines represent absence.

3.2.2 Data Sources

We analyzed repeated cross-sectional data from four waves of the UK Food and You Survey (2012–2018) [83]. This biannual survey utilizes a nationally representative sample of adults in England, Wales, and Northern Ireland. The initial sample included 11,889 participants. Following the exclusion of individuals with missing data on sociodemographics ($n = 2892$), readability perceptions ($n = 33$), or consumption ($n = 16$), the final analytical sample comprised 8948 individuals. Survey weights were applied to ensure national representativeness.

3.2.3 Outcome Variables

Participants reported consumption frequency for 14 food items. We focused on the target REMs (sandwiches, pre-cooked meats) and controls (dairy, fresh meat). The original eight-point frequency scale was collapsed into a three-point ordinal variable to ensure adequate cell counts and conceptual clarity: (1) Never; (2) Monthly aggregating once a fortnight, once a month, and less than once a month; and (3) Daily/Weekly aggregating frequencies from once or twice a week to at least once a day. This aggregation preserves the ordinal structure while producing statistically balanced groups.

3.2.4 Independent Variables

The primary independent variable is perceived MTL print size. In all waves, participants were asked: How easy do you find it to read the labelling on food products (e.g., ingredients, nutrition or storage information) in terms of the size of the print (using glasses or contact lenses if you wear them)? Responses were recorded on a 5-point Likert scale ranging from (1) Very difficult to (5) Very easy.

3.2.5 Control Variables

In addition to food consumption frequency and perceptions of MTL print size, we added several behavioural, location, and sociodemographic variables to the analysis, acknowledging their influence on REM consumption patterns [84]. This includes the presence of children under the age of 16 in the household, represented as a dichotomous variable (1 for presence, 0 for absence). Households with children tend to place a greater emphasis on nutrition, often leading to increased consumption of fruits and vegetables and reduced intake of ultra-processed foods due to heightened awareness of nutritional information [85, 86]. Studies have demonstrated that family composition, in terms of the number of members and their relationships, is associated with lower dietary diversity [87]. In this study, household size is represented by an ordinal variable encompassing family members from 1 to 4+, and marital status is a categorical variable with binary coding (1 for married or living with partner; and 0 for single, widowed, divorced, separated, or other). Research indicates that religious affiliation can impact dietary choices, with studies suggesting that individuals who are religious tend to have less diverse diets or avoid certain foods [87]. Religion is included as a dummy variable to denote Christian affiliation and other religions, with no religion serving as the reference category.

Shopping responsibilities, which indicate whether an individual is responsible for (1) all or most of the food/grocery shopping, or (0) less than half of the purchases, capture the frequency with which individuals engage in food purchasing at home. Studies suggest a decline in food purchasing and cooking at home, coinciding with the increased availability of REM in markets. Consequently, individuals who are more actively involved in home food shopping and cooking are less inclined to consume ultra-processed foods [88].

Individual characteristics such as sex and household income were also included in the analysis [89]. Sex is a binary variable, with 1 representing male and 0 representing female. Age is an ordinal variable divided into seven distinct ranges: 16–24 years, 25–34 years, 35–44 years, 45–54 years, 55–64 years, 65–74 years, and 75 years and above. Household

income was originally recorded in four categorical bands and dichotomized into a binary variable (1 = high, 0 = low) at the Â£26,000 threshold. This cutoff aligns with the original category boundaries and approximates the UK median household disposable income during the study period, effectively dividing the sample into lower- and higher-income segments.

The country of residence is also included as a dummy variable representing England, Wales, with Northern Ireland as the reference category. Another included geographical variable is the classification of urban (1) versus rural (0) areas. Recent research has indicated that diet-related concerns, such as food insecurity and obesity, are related to area-level deprivation in the UK [90]. Area-level deprivation indices were omitted because they were inconsistently available across the survey waves analyzed.

Although biological samples (e.g., urine, blood) were collected from a subset of participants in specific waves, these were not utilized in the current study. These samples remain a distinct avenue for future research, specifically to explore biomarkers related to the impact of MTL labelling on cardiovascular health.

3.2.6 Subgroup Analysis

To explore the salience-to-understanding pathway, we stratified analyses by consumer engagement [21]. Food safety concern was measured by the statement *I often worry about whether the food I have is safe to eat.* The original 5-point Likert scale was dichotomised into a high concern group (comprising those who selected definitely agree or tend to agree) and a low concern group (comprising all other respondents). Information-seeking behavior classified respondents as seekers or non-seekers based on their reported sources of food safety information. These moderators were selected as proxies for sustained label engagement, prioritizing persistent traits over acute experiences (e.g., food poisoning). Measures of visual acuity or validated health literacy scales were not available across all waves.

3.2.7 Statistical Analysis

We employed cumulative link proportional odds ordinal logistic regressions to investigate the associations between individuals' perceptions of MTL print size and the consumption frequency of the target products (pre-packaged sandwiches and pre-cooked meat) and the non-equivalent dependent variables (dairy and fresh meat). The model is specified as follows:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \alpha_j - \beta_1 \text{MTL} - \beta_2(\text{MTL} \times \text{Years}) - \boldsymbol{\gamma}\mathbf{C}$$

where Y denotes the ordinal result of the frequency of consumption for the specific food category analysed. j indexes the thresholds between the result categories. The threshold-specific intercepts α_j define the baseline log-odds of being below each threshold. β_1 represents the coefficient for the perception of MTL print size, and β_2 represents the interaction between readability and Years. $\mathbf{C}$ represents the vector of all control covariates (including the main effect of Years, sociodemographic factors, and behavioural characteristics), and $\boldsymbol{\gamma}$ is the corresponding vector of regression coefficients for these controls. A negative coefficient for β_1 indicates that participants who find the MTL print size easy to read are less likely to fall into a higher consumption frequency category.

Survey year was modeled as a categorical rather than continuous variable (with 2012 serving as the reference category). This specification enables detection of non-linear temporal trends and policy-specific period effects. Multicollinearity was assessed using Variance Inflation Factors (VIF) (see Appendixes Section A.1 and Section A.1.1). All predictors demonstrated low collinearity, with adjusted VIF values ranging from 1.02 to 1.65. These values are well below the conventional threshold of 5, confirming that the regression estimates are not biased by collinearity.

We tested the proportional odds assumption using the Brant test (see Appendixes Section A.1 and Section A.1.1). While the global test was statistically significant ($p < 0.001$)—a common outcome in large samples driven by minor deviations in de-

mographic covariates—the assumption of parallel regression held for the readability variable across all four product models (all $p > 0.50$). Based on the results of this test, we utilized ordinal logistic regression models with proportional odds.

The analyses were carried out using the statistical software R version 4.4.2 [91], with a significance level set at 5%.

3.3 Results

3.3.1 Participant Characteristics

Table Table 3.1 details the sociodemographic characteristics of the 8948 participants, stratified by their consumption frequency of pre-packaged sandwiches, pre-cooked meat, dairy, and fresh meat. The demographic profiles of these consumer segments illustrate distinct consumption patterns across product categories. For sandwiches, high-frequency consumption was notably more prevalent among women (62.4%) compared to non-consumers (45.8%). This segment of frequent consumers was also significantly younger, with individuals in the 18–24 (19.9%) and 25–34 (23.8%) age brackets being highly represented, whereas the 65+ age group predominantly comprised non-consumers (27.8%). Furthermore, individuals without religious affiliation demonstrated a greater concentration among weekly consumers (41.9%) compared to those who never consumed these products (29.3%).

In contrast, the demographic profile of pre-cooked meat consumers presents distinct characteristics. Specifically, the gender distribution transitions from a predominantly male (55.0%) representation among non-consumers to a majority female (53.0%) composition among those who consume these products weekly. Contrary to sandwich consumption patterns, older age cohorts exhibit higher consumption rates, with individuals aged 65+ forming a significant proportion of both monthly (24.7%) and weekly (20.7%) consumers. Furthermore, religious affiliation demonstrates a divergent trend, as Christians constitute

a substantial majority of weekly consumers (63.8%), whereas individuals identifying with other religions are disproportionately prevalent among non-consumers (22.8%). The consumption patterns for dairy and fresh meat also reveal unique demographic distributions. Dairy product consumption is nearly universal, with the predominant portion of the sample belonging to the weekly consumption category. This segment demonstrates a balanced gender representation (49.9% male) and comprises a considerable proportion of older consumers (21.0% aged 65+). The demographic profile for fresh meat consumption largely corresponds to that of pre-cooked meat, indicating a shift from a majority male (59.4%) non-consumer group to a majority female (53.1%) weekly consumer group, alongside a comparably high incidence of older adults (21.8%) within the weekly consumption cohort.

Across all four product categories, lower-income households consistently comprised a larger proportion of monthly and weekly consumer segments compared to non-consumers. Similarly, characteristics such as household size and marital status exhibited heterogeneous distributions across consumption levels, thereby highlighting the distinct demographic compositions of each consumer segment.

Table 3.1. Sociodemographic characteristics of participants by food consumption frequency (% within each consumption level).

	Pre-Packaged Sandwich			Pre-Cooked Meat			Dairy			Fresh Meat		
Category	**Never**	**Monthly**	**Weekly**	**Never**	**Monthly**	**Weekly**	**Never**	**Monthly**	**Weekly**	**Never**	**Monthly**	**Weekly**
n = 8948	**4052**	**3221**	**1675**	**1615**	**2035**	**5298**	**246**	**203**	**8498**	**789**	**2197**	**5963**
Sex												
Male	54.2	51.3	37.6	55.0	54.0	47.0	53.6	52.9	49.9	59.4	55.1	46.9
Female	45.8	48.7	62.4	45.0	46.0	53.0	46.4	47.1	50.1	40.6	44.9	53.1
Age												
18-24	7.7	10.9	19.9	14.6	9.0	10.9	8.1	14.7	11.1	12.5	10.7	11.1
25-34	14.2	18.1	23.8	19.3	18.4	16.4	18.1	21.1	17.3	17.4	17.8	17.2
35-44	15.5	18.1	19.7	18.2	15.4	17.6	18.5	15.7	17.2	19.1	17.7	16.8
45-54	19.1	20.3	20.0	20.8	17.8	20.1	19.3	20.6	19.7	20.2	21.0	19.2
55-64	15.7	14.3	8.2	11.2	14.7	14.2	19.2	12.5	13.7	11.7	14.2	13.9
65+	27.8	18.4	8.4	15.9	24.7	20.7	16.8	15.5	21.0	19.1	18.6	21.8
Religion												
Christian	63.8	58.0	48.7	42.1	59.5	63.8	55.5	54.0	59.1	42.4	57.4	61.6
Other religion	6.9	6.3	9.4	22.8	6.2	2.8	8.9	6.8	7.1	15.9	6.4	6.3
No religion	29.3	35.7	41.9	35.2	34.3	33.5	35.6	39.1	33.8	41.7	36.2	32.1
Marital status												
Married	39.0	37.5	51.7	44.3	41.2	39.6	46.3	51.8	40.4	45.4	42.4	39.6
Single	61.0	62.5	48.3	55.7	58.8	60.4	53.7	48.2	59.6	54.6	57.6	60.4
Household size												
1	19.5	14.2	17.9	16.5	20.1	16.5	25.3	24.7	16.9	20.5	19.6	16.0
2	38.7	39.3	28.7	30.8	41.4	37.3	30.8	29.7	37.4	29.5	36.2	38.3
3	17.4	18.3	24.2	21.2	17.9	18.7	17.6	18.3	19.0	21.8	17.7	19.1
4	24.4	28.2	29.2	31.6	20.7	27.5	26.3	27.3	26.7	28.3	26.5	26.6
Children at home												
Yes	72.6	68.8	67.7	64.4	75.8	70.0	75.3	69.3	70.2	65.6	69.8	71.1
No	27.4	31.2	32.3	35.6	24.2	30.0	24.7	30.7	29.8	34.4	30.2	28.9
Income												
Low	53.8	66.6	68.9	57.7	62.5	61.8	53.8	59.2	61.5	59.8	59.9	61.9
High	46.2	33.4	31.1	42.3	37.5	38.2	46.2	40.8	38.5	40.2	40.1	38.1
Area												
Urban	17.2	19.1	12.3	13.4	19.0	17.3	13.8	9.6	17.2	14.4	16.2	17.6
Rural	82.8	80.9	87.7	86.6	81.0	82.7	86.2	90.4	82.8	85.6	83.8	82.4
Country												
England	91.1	92.2	93.8	93.7	92.9	91.1	94.0	92.5	91.9	93.8	93.1	91.3
Wales	5.4	5.2	4.2	4.2	4.9	5.4	3.0	4.5	5.2	4.5	4.7	5.3
Northern Ireland	3.5	2.6	2.0	2.0	2.2	3.4	3.0	3.0	2.9	1.7	2.1	3.3

Table Table 3.2 indicates differentiated behavioral profiles across the product categories. Specifically, a discernible pattern regarding shopping responsibility is evident among sandwich consumers, where individuals with reduced responsibility (Sometimes/Never) exhibit a higher concentration within the frequent consumption segment (56.6

Concerns pertaining to safe eating practices exhibited divergent relationships across the product categories. For pre-packaged sandwiches, a higher prevalence of individuals with significant safety concerns was observed among weekly (25.8

Temporal analysis of the survey year variable indicates shifts in the consumer base across all product categories. For sandwich consumption, the 2018 cohort demonstrated the highest proportion among monthly consumers (28.6

Table 3.2. Behavioral characteristics of participants by food consumption frequency (% within each consumption level).

	Pre-Packaged Sandwich			**Pre-Cooked Meat**			**Dairy**			**Fresh Meat**		
Category	**Never**	**Monthly**	**Weekly**	**Never**	**Monthly**	**Weekly**	**Never**	**Monthly**	**Weekly**	**Never**	**Monthly**	**Weekly**
n = 8948	**4052**	**3221**	**1675**	**1615**	**2035**	**5298**	**246**	**203**	**8498**	**789**	**2197**	**5963**
Shopping responsability												
All/Most	56.3	50.9	43.4	51.7	56.2	50.4	58.9	53.6	51.7	53.1	54.3	50.9
Sometimes/Never	43.7	49.1	56.6	48.3	43.8	49.6	41.1	46.4	48.3	46.9	45.7	49.1
Concerns safe eating												
High	22.2	17.7	25.8	24.8	20.3	20.5	34.3	24.9	20.8	21.1	21.3	21.2
Low	77.8	82.3	74.2	75.2	79.7	79.5	65.7	75.1	79.2	78.9	78.7	78.8
Information seeking												
Yes	79.5	83.7	84.1	81.4	82.4	81.8	83.4	80.2	81.9	81.7	82.8	81.5
No	20.5	16.3	15.9	18.6	17.6	18.2	16.6	19.8	18.1	18.3	17.2	18.5
Year												
2012	26.0	22.1	22.0	20.7	18.8	26.8	15.3	13.9	24.3	18.1	17.5	27.0
2014	28.2	23.0	27.1	23.8	21.9	28.4	21.8	28.2	26.2	20.4	20.5	28.9
2016	25.8	26.4	24.8	27.7	29.1	24.0	26.2	17.6	26.0	32.1	28.2	24.1
2018	20.1	28.6	26.0	27.8	30.2	20.8	36.7	40.3	23.5	29.4	33.7	20.1

3.3.2 Perceived MTL Print Size Readability

Figure Fig. 3.4 illustrates the perceived readability of MTL print size among diverse demographic and behavioral subgroups from 2012 to 2018. Where higher scores denote enhanced readability, the findings delineate significant disparities in food label perception across distinct subgroups. A pronounced and consistent age-related pattern is observed. Specifically, younger respondents (16–24 years) consistently reported the greatest ease of label comprehension, with their mean scores demonstrating an upward trajectory from 3.91 in 2012 to a peak of 4.23 in 2018. In contrast, middle-aged cohorts (45–54) experienced the most significant challenges in readability, indicated by a decline in scores from 2.88 to 2.58 during the corresponding timeframe. Older adults (65+), however, reported a moderate level of reading ease, characterized by stable mean scores approximating 3.2.

Household and socioeconomic factors also reveal distinct patterns. Respondents from high-income households (means around 3.5–3.6) consistently found labels easier to read than those from low-income households (means around 3.1–3.3). Similarly, those with no children at home reported significantly greater ease of reading (mean 3.75) than those with children (mean approximately 3.3). Geographically, residents in Wales reported finding labels easiest to read in 2012 (mean = 3.92), though this level later converged with England and Northern Ireland. Behavioral characteristics were also associated with perceived readability. Individuals who do not actively seek food information (mean 3.50 in 2012) reported greater ease of reading than active information seekers (mean 3.33 in 2012). Likewise, those with less shopping responsibility (Sometimes/Never) reported finding labels easier to read (mean approximately 3.5) than those with primary responsibility (mean around 3.4). Differences based on food safety concerns were less pronounced, with both high- and low-concern groups reporting similar levels of readability.

Regarding pre-packaged sandwiches, a distinct and consistent hierarchical pattern in perceived readability is evident across various consumption subgroups. Specifically, individuals categorized as never consumers consistently reported the highest mean readability

scores, fluctuating between 3.78 in 2012 and 3.66 in 2018. Conversely, weekly consumers consistently indicated the lowest mean readability. In the context of pre-cooked meat, the observed pattern is less clearly delineated. Although monthly consumers generally reported the greatest ease of reading (mean 3.63 in 2012), the readability values across all three consumption segments were considerably similar and exhibited convergence over the study period, thus precluding the establishment of an enduring hierarchy. Dairy products revealed a hierarchical structure akin to that observed for sandwiches, albeit less pronounced: never consumers (mean 3.47 in 2012) and monthly consumers (mean 3.44 in 2012) consistently reported higher readability levels compared to the predominant weekly consumer subgroup. For fresh meat, the pattern is again less distinct and more closely resembles pre-cooked meat.

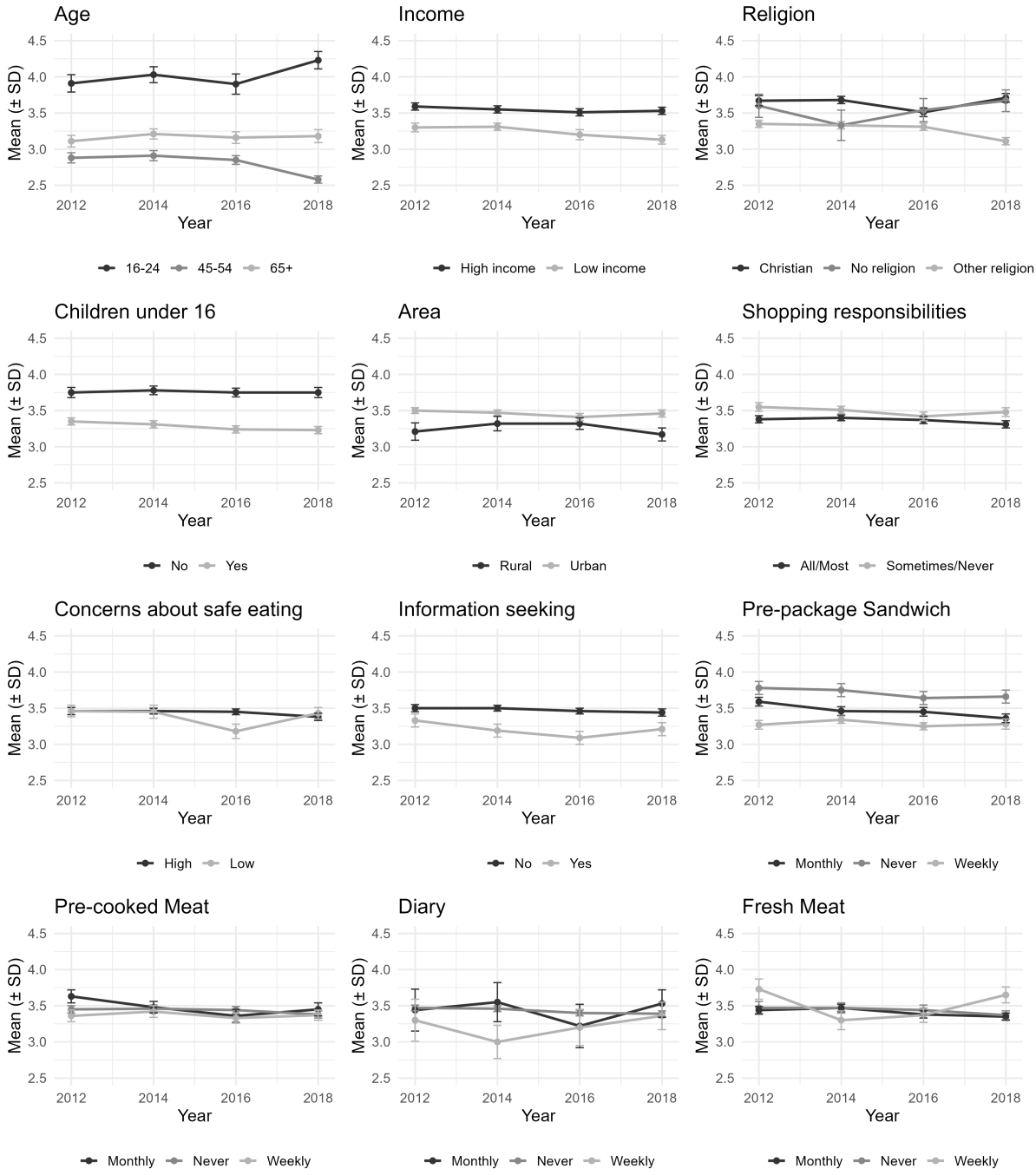

Fig. 3.4. Mean perceived MTL print size readability from 2012 to 2018, stratified by sociodemographic characteristics, behavioral characteristics, food products. Higher scores denote enhanced readability.

3.3.3 Perceived MTL Print Size Readability and Food Consumption Associations

Table Table 3.3 and Figure Fig. 3.5 present the findings derived from the ordinal logistic regressions conducted using non-equivalent dependent variables (NEDVs). Initial analysis revealed a statistically significant association exclusively concerning pre-packaged sandwiches ($\beta = -0.10, p < 0.05$). Estimations elucidate a more intricate relationship when considering the interaction between perceived readability and the survey year. Specifically for pre-packaged sandwiches (Panel A), a notable dynamic relationship was discerned. During the baseline year of 2012, a statistically significant inverse relationship was observed ($OR = 0.91$, 95% CI [0.83, 0.99]), where in each one-unit increment in perceived readability correlated with an approximate 9% reduction in the odds of more frequent consumption. This negative relationship, reflected in the interaction terms, persisted through 2014 (-4%) and 2016 (-5%). This association underwent a transformation over the study period, evidenced by a significant positive interaction term for 2018 (Readability $\times$ 2018: $OR = 1.15$, 95% CI [1.02, 1.29]). Consequently, by 2018, the overall impact of readability had inverted (combined $OR = 1.04$), manifesting as a 4% increase in the odds of frequent consumption for each one-unit rise in perceived readability.

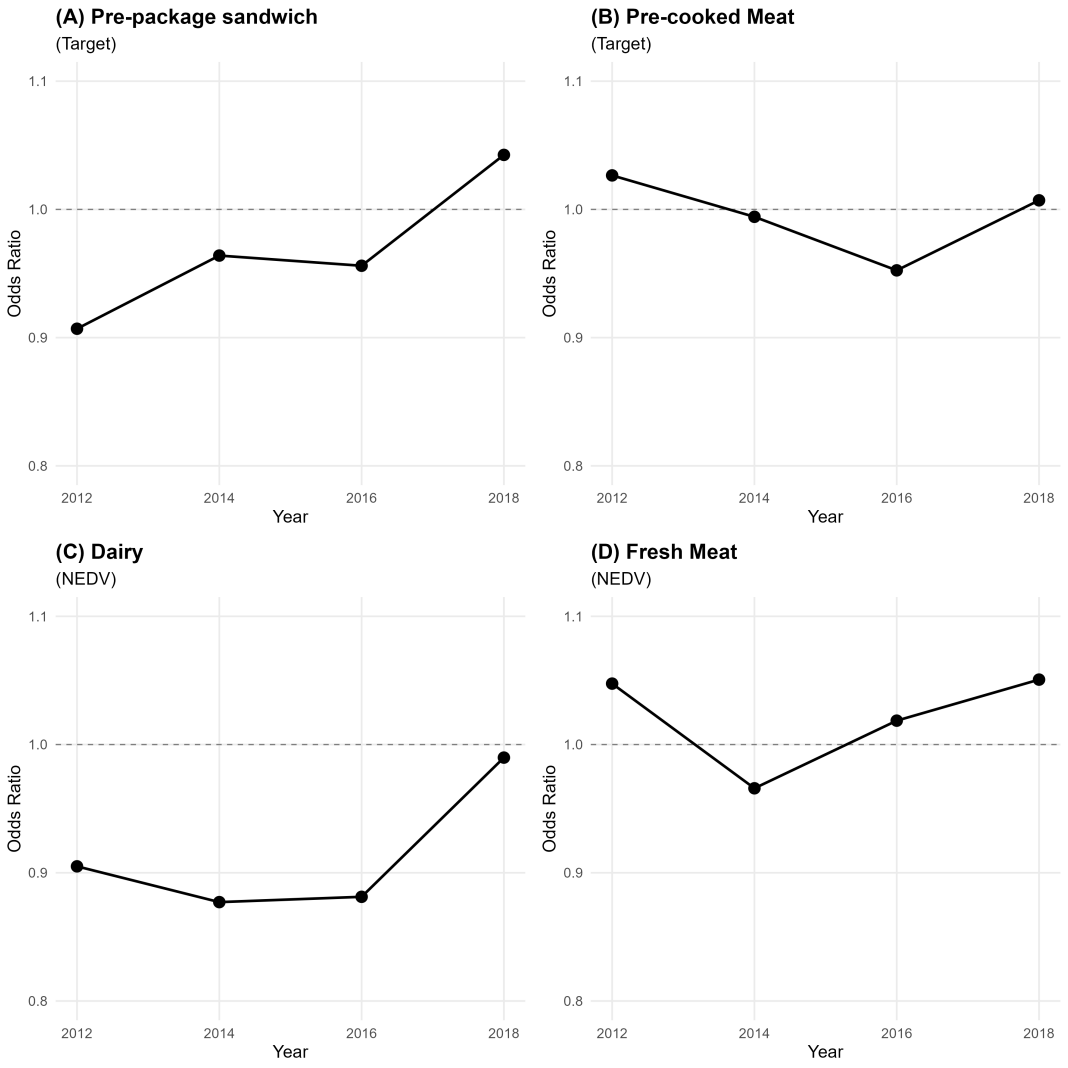

Fig. 3.5. Perceived MTL print size readability and consumption frequency. Cross-sectional trends by product type.

Table 3.3. Ordinal logistic regression results: associations of Perceived MTL print size readability and food consumption frequency, including interactions (adjusted for all sociodemographic, behavioural, and temporal variables).

	Pre-Packaged Sandwich		**Pre-Cooked Meat**		**Dairy**		**Fresh Meat**	
Variable	***β* (SE)**	**OR [95% CI]**	***β* (SE)**	**OR [95% CI]**	***β* (SE)**	**OR [95% CI]**	***β* (SE)**	**OR [95% CI]**
Readability	−0.10 * (0.04)	0.91 [0.83, 0.99]	0.03 (0.03)	1.03 [0.96, 1.10]	−0.10 (0.11)	0.90 [0.72, 1.13]	0.05 (0.04)	1.05 [0.96, 1.14]
Year (vs. 2012)								
2014	−0.28 (0.23)	0.76 [0.49, 1.18]	0.15 (0.21)	1.16 [0.77, 1.75]	0.55 (0.60)	1.73 [0.53, 5.63]	0.27 (0.25)	1.31 [0.81, 2.12]
2016	−0.39 * (0.22)	0.68 [0.44, 1.05]	0.63 *** (0.19)	1.87 [1.30, 2.70]	0.52 (0.54)	1.68 [0.59, 4.82]	0.72 *** (0.20)	2.06 [1.39, 3.05]
2018	−0.96 *** (0.21)	0.38 [0.25, 0.58]	0.55 ** (0.19)	1.73 [1.20, 2.49]	0.78 (0.50)	2.19 [0.82, 5.84]	0.84 *** (0.20)	2.31 [1.57, 3.40]
Interactions								
Readability × 2014	0.06 (0.06)	1.06 [0.94, 1.20]	−0.03 (0.06)	0.97 [0.86, 1.09]	−0.03 (0.17)	0.97 [0.69, 1.35]	−0.08 (0.07)	0.92 [0.80, 1.06]
Readability × 2016	0.05 (0.06)	1.05 [0.93, 1.19]	−0.07 (0.05)	0.93 [0.84, 1.03]	−0.03 (0.16)	0.97 [0.72, 1.32]	−0.03 (0.06)	0.97 [0.87, 1.09]
Readability × 2018	0.14 * (0.06)	1.15 [1.02, 1.29]	−0.02 (0.05)	0.98 [0.89, 1.08]	0.09 (0.14)	1.09 [0.83, 1.44]	0.00 (0.06)	1.00 [0.89, 1.12]

Note: Table displays coefficients (β) with standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This dynamic association was not observed across the control outcome categories. Specifically, the analysis for pre-cooked meat (Panel B), dairy (Panel C), and fresh meat (Panel D) revealed no comparable temporal dynamic; neither the main association with readability nor its interactions with the survey year achieved statistical significance. This consistent pattern robustly indicates that the observed dynamic association is not a generalised consequence of label salience but rather an effect specific to the pre-packaged sandwich category. While this dynamic interaction was specific to this category, other significant temporal trends were observed across all models. Specifically, for pre-packaged sandwiches, the odds of frequent consumption in 2018 were reduced by 62% compared to those in 2012. In contrast, the odds of frequent consumption for pre-cooked meat demonstrated a significant elevation in subsequent years, increasing by 87% in 2016 and 73% in 2018 relative to the 2012 baseline. Furthermore, fresh meat consumption exhibited an even more pronounced increase, with the odds of frequent consumption in 2018 exceeding a two-fold rise.

3.3.4 Health Belief Associations

Tables Table 3.4 and Table 3.5 show the findings from ordinal logistic regressions, stratified by health belief behavioral subgroups. The results confirms that the relationship between perceived MTL print size readability and consumption frequency is significantly influenced by product category, suggesting varied patterns of label engagement.

The analysis stratified by information-seeking behavior highlights a key divergence. Among non-information seekers, a significant negative correlation was observed between perceived readability and the frequency of pre-packaged sandwich consumption ($OR = 0.81$, 95% CI [0.69, 0.95]). In other words, each one-unit increment in perceived MTL print size readability correlated with an approximate 19% reduction in the odds of more frequent consumption. While the primary association did not achieve significance among information seekers, a statistically significant interaction with the year 2018 was

exclusively observed within this group ($OR = 1.15$, 95% CI [1.01, 1.30]). This indicates that their relationship with readability has evolved, manifesting as a 7% increase in the odds of frequent consumption for each one-unit rise in perceived readability.

For the control pre-cooked meat, readability consistently showed no significant main or interaction effects in either group. This null finding is mirrored in the no-label falsification test, fresh meat. Notably, the dairy control analysis revealed a pattern analogous to that of sandwiches, where a significant main effect of readability was exclusively identified among non-information-seeking participants ($OR = 0.65$, 95% CI [0.46, 0.92]).

When stratifying by concerns regarding food safety, a comparable pattern was observed. Specifically for pre-packaged sandwiches, both the significant primary association with readability ($OR = 0.90$, 95% CI [0.82, 0.99]) and the notable interaction effect with the year 2018 ($OR = 1.16$, 95% CI [1.03, 1.31]) were observed solely among individuals expressing low levels of concern. In this subgroup, a one-unit increment in perceived MTL print size readability initially correlated with an approximate 10% reduction in the odds of more frequent consumption, a trend that subsequently evolved into a 4% increase by 2018. In the control categories of pre-cooked meat and dairy, readability did not emerge as a significant factor for either the high- or low-concern subgroups. Conversely, the no-label fresh meat model unexpectedly revealed a significant positive interaction with the year 2018 specifically within the high-concern group ($OR = 1.40$, 95% CI [1.08, 1.81]), despite the main effect of readability not reaching statistical significance for either subgroup.

3.3.5 Sensitivity Analysis

As a robustness check, we reframed the analysis using binary logistic regression, where consumption was classified as either Yes (any frequency) or No (never), while adjusting for the same control variables (see Appendixes Section A.1 and Section A.1.1). The results from this binary approach reveal a consistent directional association for pre-packaged sandwiches compared to the ordinal model. For this product, a statistically significant

Table 3.4. Ordinal logistic regression results: associations of perceived MTL print size readability and food consumption frequency by information-seeking subgroups (adjusted for all sociodemographic, behavioural, and temporal variables).

	Information Seekers		**Non-Information Seekers**	
Variable	**β (SE)**	**OR [95% CI]**	**β (SE)**	**OR [95% CI]**
Pre-packaged Sandwich				
Readability	−0.07 (0.05)	0.94 [0.85, 1.03]	−0.22 ** (0.08)	0.81 [0.69, 0.95]
Year (vs. 2012)				
2014	−0.17 (0.25)	0.84 [0.52, 1.36]	−0.47 (0.47)	0.63 [0.25, 1.57]
2016	−0.41 * (0.24)	0.66 [0.41, 1.06]	−0.28 (0.45)	0.75 [0.31, 1.81]
2018	−0.90 *** (0.23)	0.41 [0.26, 0.64]	−1.10 ** (0.43)	0.33 [0.14, 0.77]
Readability × Year				
2014 vs. 2012	0.05 (0.07)	1.05 [0.91, 1.20]	0.04 (0.13)	1.04 [0.80, 1.34]
2016 vs. 2012	0.08 (0.07)	1.08 [0.94, 1.24]	−0.08 (0.12)	0.93 [0.73, 1.17]
2018 vs. 2012	0.14 * (0.06)	1.15 [1.01, 1.30]	0.12 (0.12)	1.12 [0.88, 1.43]
Pre-cooked Meat				
Readability	0.06 (0.05)	1.06 [0.97, 1.16]	−0.06 (0.08)	0.94 [0.79, 1.11]
Year (vs. 2012)				
2014	0.24 (0.26)	1.28 [0.77, 2.12]	−0.36 (0.46)	0.70 [0.28, 1.73]
2016	0.66 ** (0.24)	1.93 [1.21, 3.09]	0.83 * (0.40)	2.29 [1.04, 5.04]
2018	0.76 *** (0.24)	2.15 [1.35, 3.42]	0.51 (0.37)	1.67 [0.80, 3.46]
Readability × Year				
2014 vs. 2012	−0.06 (0.07)	0.95 [0.82, 1.09]	0.12 (0.13)	1.13 [0.87, 1.46]
2016 vs. 2012	−0.06 (0.07)	0.94 [0.82, 1.07]	−0.16 (0.13)	0.85 [0.67, 1.09]
2018 vs. 2012	−0.05 (0.07)	0.95 [0.84, 1.08]	−0.06 (0.11)	0.95 [0.76, 1.18]
Dairy				
Readability	−0.02 (0.13)	0.98 [0.75, 1.27]	−0.44 * (0.18)	0.65 [0.46, 0.92]
Year (vs. 2012)				
2014	0.61 (0.73)	1.85 [0.44, 7.74]	0.55 (0.81)	1.74 [0.36, 8.44]
2016	0.71 (0.64)	2.03 [0.58, 7.12]	−0.02 (0.90)	0.98 [0.17, 5.67]
2018	1.22 * (0.57)	3.40 [1.11, 10.4]	−0.82 (0.81)	0.44 [0.09, 2.16]
Readability × Year				
2014 vs. 2012	−0.06 (0.20)	0.94 [0.63, 1.39]	0.06 (0.25)	1.06 [0.65, 1.73]
2016 vs. 2012	−0.07 (0.18)	0.93 [0.65, 1.33]	0.15 (0.29)	1.16 [0.66, 2.05]
2018 vs. 2012	0.00 (0.16)	1.00 [0.73, 1.37]	0.45 * (0.25)	1.57 [0.97, 2.55]
Fresh Meat				
Readability	0.09 (0.06)	1.09 [0.98, 1.22]	−0.08 (0.10)	0.92 [0.75, 1.13]
Year (vs. 2012)				
2014	0.53 (0.28)	1.69 [0.98, 2.92]	−0.19 (0.52)	0.82 [0.30, 2.26]
2016	0.90 *** (0.26)	2.45 [1.47, 4.09]	0.44 (0.47)	1.55 [0.62, 3.87]
2018	1.09 *** (0.25)	2.97 [1.82, 4.84]	0.30 (0.44)	1.35 [0.57, 3.20]
Readability × Year				
2014 vs. 2012	−0.13 (0.08)	0.88 [0.75, 1.03]	0.03 (0.14)	1.03 [0.79, 1.36]
2016 vs. 2012	−0.06 (0.07)	0.94 [0.82, 1.09]	0.03 (0.14)	1.03 [0.78, 1.35]
2018 vs. 2012	−0.05 (0.07)	0.95 [0.83, 1.09]	0.14 (0.13)	1.15 [0.89, 1.48]

Note: Table displays coefficients (β), standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3.5. Ordinal logistic regression results: associations of perceived MTL print size readability and food consumption frequency by concern subgroups (adjusted for all sociodemographic, behavioural, and temporal variables).

	High Concern		**Low Concern**	
Variable	**β (SE)**	**OR [95% CI]**	**β (SE)**	**OR [95% CI]**
Pre-packaged Sandwich				
Readability	−0.09 (0.10)	0.92 [0.76, 1.12]	−0.11 * (0.05)	0.90 [0.82, 0.99]
Year (vs. 2012)				
2014	−0.48 (0.47)	0.62 [0.24, 1.55]	−0.17 (0.24)	0.84 [0.53, 1.34]
2016	−0.66 (0.48)	0.52 [0.20, 1.32]	−0.29 (0.24)	0.75 [0.47, 1.20]
2018	−0.84 (0.46)	0.43 [0.18, 1.06]	−1.00 *** (0.24)	0.37 [0.23, 0.58]
Readability × Year				
2014 vs. 2012	0.12 (0.13)	1.13 [0.87, 1.46]	0.03 (0.07)	1.03 [0.91, 1.17]
2016 vs. 2012	0.11 (0.13)	1.12 [0.86, 1.44]	0.03 (0.07)	1.03 [0.90, 1.18]
2018 vs. 2012	0.11 (0.13)	1.12 [0.87, 1.44]	0.15 * (0.06)	1.16 [1.03, 1.31]
Pre-cooked Meat				
Readability	0.04 (0.09)	1.04 [0.88, 1.23]	0.02 (0.05)	1.02 [0.93, 1.12]
Year (vs. 2012)				
2014	0.05 (0.48)	1.06 [0.41, 2.73]	0.18 (0.26)	1.19 [0.71, 2.00]
2016	0.85 * (0.43)	2.35 [1.01, 5.46]	0.63 ** (0.23)	1.87 [1.20, 2.92]
2018	0.69 (0.45)	2.00 [0.83, 4.81]	0.64 ** (0.22)	1.90 [1.24, 2.92]
Readability × Year				
2014 vs. 2012	0.05 (0.13)	1.05 [0.81, 1.36]	−0.05 (0.07)	0.95 [0.83, 1.09]
2016 vs. 2012	−0.17 (0.12)	0.85 [0.67, 1.07]	−0.05 (0.06)	0.95 [0.84, 1.08]
2018 vs. 2012	0.01 (0.13)	1.01 [0.79, 1.30]	−0.05 (0.06)	0.96 [0.85, 1.08]
Dairy				
Readability	−0.04 (0.17)	0.96 [0.69, 1.34]	−0.12 (0.15)	0.89 [0.66, 1.19]
Year (vs. 2012)				
2014	1.32 (0.90)	3.76 [0.64, 22.0]	0.22 (0.74)	1.25 [0.29, 5.37]
2016	0.29 (0.89)	1.34 [0.24, 7.61]	0.59 (0.69)	1.80 [0.47, 6.94]
2018	1.24 * (0.71)	3.46 [0.85, 14.0]	0.63 (0.65)	1.87 [0.52, 6.70]
Readability × Year				
2014 vs. 2012	−0.25 (0.26)	0.78 [0.47, 1.29]	0.05 (0.21)	1.05 [0.70, 1.59]
2016 vs. 2012	0.09 (0.26)	1.10 [0.66, 1.82]	−0.06 (0.19)	0.94 [0.65, 1.37]
2018 vs. 2012	−0.04 (0.22)	0.96 [0.63, 1.47]	0.15 (0.18)	1.16 [0.81, 1.65]
Fresh Meat				
Readability	−0.14 (0.10)	0.87 [0.72, 1.06]	0.11 (0.05)	1.11 [0.99, 1.24]
Year (vs. 2012)				
2014	−0.47 (0.51)	0.63 [0.23, 1.70]	0.63 * (0.28)	1.88 [1.09, 3.24]
2016	0.07 (0.43)	1.07 [0.47, 2.47]	1.01 *** (0.25)	2.74 [1.68, 4.45]
2018	−0.34 (0.45)	0.71 [0.29, 1.71]	1.23 *** (0.24)	3.44 [2.16, 5.46]
Readability × Year				
2014 vs. 2012	0.16 (0.15)	1.17 [0.87, 1.57]	−0.17 * (0.08)	0.84 [0.72, 0.99]
2016 vs. 2012	0.14 (0.12)	1.15 [0.90, 1.47]	−0.09 (0.07)	0.92 [0.80, 1.05]
2018 vs. 2012	0.34 ** (0.13)	1.40 [1.08, 1.81]	−0.09 (0.07)	0.91 [0.80, 1.04]

Note: Table displays coefficients (β), standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

negative association emerged suggesting that each one-unit increment in perceived MTL print size readability correlated with an approximate 16% reduction in the odds of being a consumer. The findings for the control outcome categories aligned with those obtained from the primary ordinal model, indicating an absence of a statistically significant association between perceived readability and the propensity for consumption. While this robustness check confirms the product-specific nature of the findings, it is important to note that the binary classification of consumption is a simplification. This approach may mask more subtle relationships by not distinguishing between different levels of consumption frequency, such as monthly versus weekly, which were considered in the primary ordinal analysis.

We conducted a sensitivity analysis to address the possibility of reverse causality. Specifically, we modelled readability as the outcome and consumption frequency as the predictor, adjusting for the same covariates (see Appendixes Section A.1 and Section A.1.1). The results indicate that the direct association between these two variables was statistically significant for both pre-packaged sandwiches and pre-cooked meat. This significant reverse association underscores the importance of carefully considering the directional implications within the primary model. While our findings support strong reverse causality, the cross-sectional nature of the data means we cannot definitively establish the direction of the relationship. Future longitudinal studies are needed to clarify these dynamics.

While ready meal constitutes a distinct category within the UK Food and You Survey, its data availability was limited to 2016, thereby precluding the analysis of cross-sectional associations central to this research. For transparency, the ordinal logistic regression model for this specific category was included in the Appendixes Section A.1 and Section A.1.1. The resulting positive, non-statistically significant estimates corroborate the findings of our primary model for pre-cooked meat, although caution is required due to the small sample size.

To validate the assumption of linearity, we conducted a sensitivity analysis treating perceived print size readability as a categorical variable rather than a continuous one. The Target Product Sandwiches is the only category where perceived readability shows a

significant negative association with consumption frequency (e.g., $OR = 0.50$ for Easy vs Very difficult), supporting the linear gradient observed in the main analysis (see Appendixes Section A.1 and Section A.1.1). In contrast, the coefficients for all three NEDVs (Pre-cooked Meat, Dairy, and Fresh Meat) are statistically non-significant ($p > 0.05$) and fluctuate without a clear linear pattern. This confirms that the observed effect is specific to the target product and robust to the functional form of the variable.

We conducted an analysis comparing the primary full model against a parsimonious model. We specifically excluded religion, household size, urban/rural status, and the presence of children under 16 from the reduced model, as these variables demonstrated limited predictive power and statistically non-significant associations with the outcome in the full model. As shown in the Appendixes Section A.1 and Section A.1.1, the coefficient for perceived readability remained stable across all product categories. For the target product (sandwiches), the coefficient changed by less than 1%, confirming that the included demographic variables did not introduce confounding bias. Consequently, the full model was retained in the main analysis to strictly adjust for all potential sociodemographic variations.

Transportability is a critical component of external validity cheques [32, 43], adjusting study results between two UK shopper populations by accounting for differences in baseline characteristic distributions [92]. In this study, our validation focused specifically on the sampling mechanism, accounting for the structural differences between the perception-based UK Food and You survey (approx. 3000 participants per wave) and the objective National Diet and Nutritional Survey (NDNS) (approx. 1000 participants per wave; 4-day consumption report).

We accomplished this by generalizing the Food and You findings to the NDNS target population, employing two generalized boosted models (GBM) with distinct sets of covariates. The first model included only demographic and socioeconomic covariates, while the second, which we refer to as the fully adjusted model in our main analysis, encompassed a broader set of variables. This approach strengthens external validity by mitigating selection biases inherent in the differing sampling designs. By demonstrating

that the association holds even when the source sample is reweighted to match the demographic structure of the NDNS, we confirm that the findings are applicable to the broader policy-relevant population and are not idiosyncratic to the Food and You cohort.

Furthermore, to evaluate how well the transportability weights from our fully adjusted model balanced the covariate distributions, we performed two diagnostic checks. First, we created a table comparing the distributions of the shared covariates between the Food and You and NDNS samples. Second, we plotted the density of the estimated participation scores for both survey samples to visually assess their overlap and the success of the weighting adjustment.

3.4 Discussion

This study investigated the association between the perceived readability of MTL labelling print size and consumer choice, utilizing a non-equivalent dependent variable (NEDV) design to isolate content-specific effects [71, 72]. Our findings reveal a relationship that is not merely product-specific but contingent on the interaction between the label content and the product consumption context. We observed a significant, dynamic association isolated exclusively to pre-packaged sandwiches, while finding persistent null effects for our other unhealthy labeled product (pre-cooked meat), our nutritionally advisable control (dairy), and our no-label falsification test (fresh meat).

This specific pattern provides strong evidence that the salience-to-understanding pathway functions as a warning mechanism. For pre-packaged sandwiches, readability was linked to a 9% decrease in the odds of frequent consumption in 2012 ($OR = 0.91$), consistent with the intended deterrent effect of the red warning. This core finding was robust to transportability analysis [43], which validated the structural generalizability of this association to the broader dietary population represented in the NDNS. This strengthens confidence that the result is not an artifact of the specific UK Food and You Survey

sample. Furthermore, the consistent absence of an effect for dairy and fresh meat supports the causal logic of the NEDV design: where the warning signal is absent—either due to green reinforcing codes or no label at all—increasing salience does not alter behavior.

The divergence between pre-packaged sandwiches and pre-cooked meat—both unhealthy, labeled REMs—warrants specific attention. We argue this stems from the distinct decision-making contexts characterizing ambivalent convenience, specifically the cognitive conflict between the immediate utility of a time-saving meal and the nutritional ambiguity regarding its healthfulness. Sandwiches are often time-pressured, discretionary, impulse-driven purchases where the health vs. convenience trade-off is salient at the moment of choice [57, 93]. In this high-variance category, where the health halo of a product can be deceptive, a readable warning can tip the balance by resolving nutritional uncertainty.

In contrast, pre-cooked meat (e.g., ham, sliced chicken) often functions as a planned meal component or staple ingredient for evening meals [94]. For such habitual purchases, consumers may rely on top-down processing goals (e.g., I need sliced chicken for lunch) that override bottom-up visual cues like label print size [95]. Unlike sandwiches, pre-cooked meats may be viewed through a stable processed heuristic, rendering granular label details less influential. This suggests that label salience is most potent when the purchase decision is malleable and susceptible to immediate visual disruption.

Regarding the temporal dynamics, the reversal of the sandwich association by 2018—where readability became associated with increased consumption—likely reflects a combination of habituation and market adaptation. First, consistent with a time decay effect, the initial visual salience of the warning label may diminish over time as consumers habituate to the signal, reducing its disruptive power [96]. Second, empirical evidence suggests that manufacturers aggressively reformulated products following the MTL implementation to mitigate the stigma of ultra-processed foods [97, 98]. Consequently, by 2018, a readable label may have signaled transparency or quality in a reformulated market, effectively neutralizing the warning effect. Theoretically, this implies a repurposing of the pathway; readability may have ceased to function as a deterrent and instead enhanced the salience

of the product's trustworthiness.

This relationship between attention and behavior also appears to be bidirectional. Our reverse causality sensitivity analysis indicated that frequent consumption predicts lower perceived readability. This suggests a potential feedback loop driven by habituation: as individuals consume these products more frequently, they may develop label blindness, engaging less with visual cues as purchasing becomes automatic. Thus, habitual consumers may perceive labels as less readable simply because they have stopped looking at them.

Finally, regarding consumer heterogeneity, our subgroup analysis aligns with the Health Belief Model. Readability predicted consumption primarily among less-engaged consumers (non-information seekers). For these individuals, who lack strong internal cues to seek nutrition info, external visual salience acts as a necessary call to action [99]. Conversely, highly engaged consumers likely employ directed attention strategies, seeking out nutritional data regardless of print size. This challenges the information overload hypothesis [67]; the issue for less-engaged groups is not too much data, but the perceptual accessibility of the primary warning signal.

Descriptively, our findings that older adults and lower-income groups report significantly lower perceived readability raise concerns about structural inequities. These perceptual disparities may signal underlying barriers related to visual health and health literacy [65, 100, 101]. If vulnerable groups physically struggle to discern the label, the policy fails them at the first hurdle of the communication pathway, exacerbating health inequalities.

Consequently, these findings suggest that a one-size-fits-all approach to labeling standards is insufficient. While increasing print size is a viable intervention, its effectiveness is bounded by packaging constraints. Policy should therefore consider adaptive standards: mandating larger, bolder warnings specifically for impulse categories where visual disruption is effective, while exploring alternative strategies—such as digital augmentation or simplified interpretative symbols—for categories where print size is constrained or less impactful.

Limitations

This investigation offers valuable insights into real-world phenomena, though several limitations warrant consideration. We have categorized these into three key areas concerning study design and causal inference, measurement and bias, and statistical constraints.

Regarding study design and causal inference, the study employs a repeated cross-sectional design, pooling data from four waves. Although this precludes tracking individual behavioral trajectories, a strength of panel data, it suits our primary aim of modeling population-level changes in associations over time. By drawing fresh samples at each time point, it avoids attrition bias typical of long-term panel studies and maintains representativeness of the UK population throughout the policy period. However, the data structure inherently constrains the ability to draw definitive causal inferences.

Given the simultaneous nationwide implementation of the MTL policy in the UK, it was not feasible to distinguish distinct treatment groups. While our NEDV design mitigates general history threats, we acknowledge the potential for unobserved confounders that vary across food categories. For instance, unmeasured factors such as taste preferences or exposure to product-specific marketing campaigns could differentially influence the association for sandwiches compared to control products, potentially biasing the observed product-specific estimates. Consequently, the study is confined to reporting statistical associations rather than establishing causal effects. Furthermore, the observed reversal in the association for pre-packaged sandwiches may stem from concurrent, unmeasured events, such as broader public health campaigns, rather than solely reflecting individual-level adaptation to the MTL. Thus, population-level trends warrant cautious interpretation given the risk of an ecological fallacy.

With respect to measurement and bias, limitations exist regarding our key variables. Our central variable, perceived readability, is a self-reported proxy for the salience-to-understanding pathway, not a direct objective measure of comprehension. Furthermore, this measure pertains to food labelling in general, rather than specifically to the MTL

panel, introducing potential measurement error if consumers struggle with ingredient lists but not the MTL. Future research should strengthen this by employing objective validations, such as eye-tracking or standardized legibility tests (e.g., ISO standards), to definitively disentangle visual acuity from perceived comprehension.

Similarly, we utilized food safety information seeking as a proxy for general engagement with nutritional information. We acknowledge that seeking information about safety (e.g., hygiene, poisoning) is conceptually distinct from seeking nutritional data. However, in the absence of a consistent nutritional seeking metric across all survey waves, this variable was selected as the most robust available indicator of active information-acquisition behavior. We assume a degree of overlap wherein vigilant consumers who actively verify safety information demonstrate a higher baseline of label engagement that likely extends to other informational cues, including nutrition.

The consumption data is also self-reported and subject to recall or social desirability biases. Crucially, these biases are likely differential across food categories, as consumers may be more prone to underreport the consumption of unhealthy items like sandwiches and pre-cooked meat compared to neutral staples like dairy, potentially attenuating the observed associations for the target products. Finally, while our study highlights the role of print size, it is methodologically challenging to fully disentangle its effect from interacting design elements, such as color contrast, layout, and typography, within an observational dataset.

Concerning statistical and data constraints, the dataset presented specific challenges, notably a significant portion of missing income data (>20% of the sample). We elected not to use multiple imputation, as the missingness was likely not random, and imputing such a large proportion of a primary socioeconomic covariate could introduce greater bias than listwise deletion in this context. Additionally, due to inconsistencies in how consumption frequency was measured across different food categories in the UK Food and You Survey, we could not extend this comparative analysis to other types of food products. Future research could address this limitation by utilizing household scanner data, which

would allow for the analysis of a broader range of products and provide objective, granular measures of purchasing behavior.

Our measure of readability is subjective, representing a self-reported perception rather than an objective test of comprehension. Future research should strengthen this measurement strategy by employing objective validations, such as eye-tracking or standardized legibility tests (e.g., ISO standards), to definitively disentangle visual acuity from perceived comprehension and subjective readability. Finally, the content-dependent inference relies on the assumption that consumers interpret red MTLs as warnings; while supported by literature, individual interpretation of color codes varies.

3.5 Conclusions

The association between perceived MTL print size readability and consumption frequency is intricate and contingent not just on product type but on the label's content and the product's context. Our primary finding, strengthened by the use of non-equivalent dependent variables, indicates that the role of label salience is highly specific. We found a significant dynamic association isolated exclusively to pre-packaged sandwiches. This relationship reversed from a 9% decrease in the odds of frequent consumption in 2012 to a 4% net increase in the odds of frequent consumption per unit of readability by 2018. It is crucial to note that this 2018 result ($OR = 1.04$) implies that higher readability was associated with more frequent consumption, a finding that contradicts the simple hypothesis that better readability of red warnings reduces intake. Conversely, a persistent null association was found for pre-cooked meat, our nutritionally advisable-labelled control (dairy), and our no-label falsification test (fresh meat). This pattern strongly supports our hypothesis that the salience-to-understanding pathway, as proxied by readability, functions as a conditional warning mechanism. Its effect appears specific to red health warnings on ambivalent convenience foods rather than a general effect of all labels. Furthermore, the large sample size provided sufficient statistical power to detect these subtle interaction

effects, particularly within subgroups. The association for sandwiches was strongest among less-engaged consumer groups, supporting a model where readability acts as the primary behavioral cue for those lacking intrinsic motivation. Theoretically, our findings challenge the efficacy of a uniform approach to label-based interventions. Policy implications suggest that standardization alone may be insufficient. Instead, public health strategies should consider adaptive standards—mandating high-salience warnings for impulse categories where visual disruption is effective, while exploring alternative formats for categories where habituation or context neutralizes the warning signal.

Publicly available datasets were analyzed in this study. The data from the Food and You Survey can be found from the UK Data Service at `https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000053` (accessed on 03 November 2023). Additional data are available on the UK Food Standards Agency (FSA) website at `https://data.food.gov.uk/catalog/datasets/3f3ad1b7-8cf3-444b-abbf-f784ea4551e1` (accessed on 28 October 2024). Data for the National Diet and Nutrition Survey can be found at `https://datacatalogue.ukdataservice.ac.uk/studies/study/6533#details` (accessed on 09 April 2024). Further guidelines and calorie reduction targets are available from Public Health England via `https://www.gov.uk/government/organisations/public-health-england` (accessed on 28 October 2024).

Chapter 4

Food Label Granularity and Working Memory

Abstract

Objective: While prior research on front-of-package labels has focused on their design, the impact of their informational granularity—or level of detail—on consumer choice remains understudied. This study aimed to investigate how the informational granularity of food calorie labels, in conjunction with consumers' working-memory capacity, influences choice behavior. Methods: A randomized controlled trial was conducted (UK adults, N = 498) in which participants completed a cereal-shopping task after being assigned to one of three conditions: coarse label, detailed label, or no label. Working-memory capacity was evaluated via a three-level n-back test. The primary outcomes measured were the average number of calories per chosen product and the probability of selecting low-calorie options. Multilevel models were employed to test both main and interaction effects, with the Benjamini–Hochberg adjustment applied to control the false discovery rate. Results: Compared with the control condition, food labels reduced mean calorie selection by 4.56%. Each one-unit increase in 3-back performance resulted in an additional 14 kcal reduction under coarse labels and 18 kcal under detailed labels. Individuals with high working memory capacity exhibited a preference for moderate-calorie products over the lowest-calorie products when detailed labels were displayed. The effects of capacity under coarse labels were less pronounced and primarily limited to avoiding the lowest-calorie option. These findings provide conceptual insights into information optimization and offer practical guidance for policymakers on designing effective FOP schemes by specifying when detail helps versus hinders consumer choice.

4.1 Introduction

High-energy-dense foods have been identified as critical factors driving the increase in obesity rates, thereby increasing the population's risk of developing noncommunicable diseases such as type 2 diabetes, cardiovascular disease, and various forms of cancer [4, 9]. This trend has led governments worldwide to explore strategies to promote healthier eating choices, with front-of-package labelling (FOP) emerging as a key public health intervention. By providing concise and simplified 'at-a-glance' assessments of the nutritional content of foods, these labels aim to promote dietary quality in two ways: 1) improving the understanding of the nutritional quality of packaged foods [7, 8, 66] and 2) driving product reformulation [102, 19]. Traditionally, the nutritional facts displayed on packaging have been the primary source of data that many nations have relied upon to guide consumers toward making healthier food choices [103]. However, nutritional facts are challenging to comprehend, and there is limited evidence to suggest that this approach effectively influences dietary behaviors in a positive manner [104, 105]. Given that FOP labels have been demonstrated to be more salient and easier to comprehend, there has been a growing emphasis on optimizing the use of FOP labels to enhance dietary quality [7, 15, 8, 16]. Recent systematic reviews incorporating experimental and observational data have shown that labelling can decrease consumer energy intake by 6.6% and increase vegetable consumption by 13.5%, helping individuals identify healthier food options [8], although the impact on purchasing intentions remains ambiguous, and the effects on overall consumption are limited [7]. From a standard economic perspective, FOP labelling can be viewed as a disclosure policy that aims to address information asymmetry [11]. Labelling introduces transparency [13] and respects consumer freedom of choice [14] without imposing stringent regulations. Research indicates that individuals tend to exhibit preference biases and misjudge the health consequences of their behaviors [11, 10]. Consequently, implementing FOP labelling is a justified way to educate individuals about potential costs that may not be fully internalized at the time of purchase [9]. By making information available, labelling aims to encourage more informed decision-making, helping consumers better

align choices with long-term health interests. More specifically, FOP labels can influence the mechanisms underlying belief and behavioral intention, as well as those associated with planning, goal setting, and the maintenance of the behavior of interest, namely, the use of FOP labels during the purchase of high-energy-dense foods [18, 68]. However, the expected effects of this policy are ambiguous, as they depend on the prior beliefs of people about the healthiness of their shopping baskets, which may be updated after the use of labelling [16, 13]. Even if individuals have perfect knowledge of the nutritional content of products, FOP labelling can still influence their choices by prompting a heuristic learning process for food decisions [8, 15]. People's understanding of labelling is central to the efficacy of this policy. As the Grunert and Mills theoretical framework suggests, for a label to be effective, it should be seen, liked, and understood [21]. The literature indicates that labels have been found to enhance individuals' ability to select products on the basis of health status when they can objectively understand the information they contain, enabling them to accurately differentiate products according to their nutritional content [6, 33, 34, 35, 36, 106]. Additionally, labels have been shown to improve product selection abilities when individuals subjectively perceive that they have understood the nutritional information presented [6, 33, 34, 37]. Demographic attributes, personal interest, health literacy levels, and label design have all been identified as factors that can influence the understanding of labels [21, 38]. In psychology, the way in which labelled information is provided is crucial [19, 11]. As individuals exhibit limited attention and cognitive resources, they face a trade-off between the time and effort required to locate and interpret information [12, 22, 23]. Therefore, a key factor is the simplicity of the information used [10]. The literature has not yet reached a definitive consensus on which FOP labelling format is the most readily legible and understandable for consumers and which consequently promotes the healthiest purchasing decisions. Existing evidence suggests that interpretative labels, which convey information about the healthiness of a food, tend to be more effective than noninterpretive systems that do not involve any judgement, such as reference intakes. Prominent examples of interpretative labels include nutrient-specific systems such as Multiple Traffic Light (MTL) and Healthy Start Rating (HSR); nutrient-specific warnings such as Keyhole

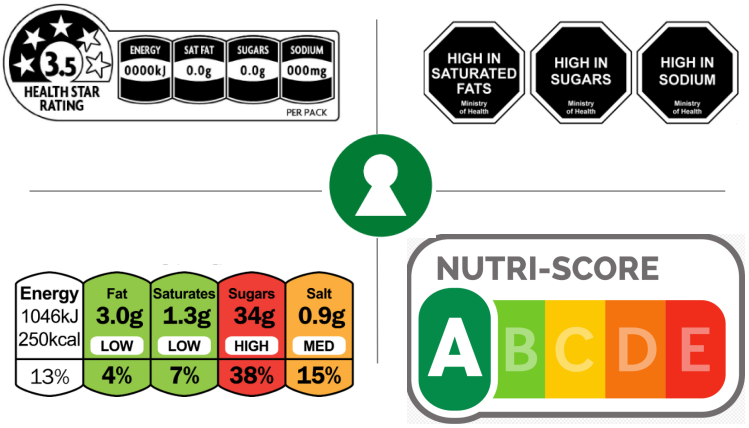

Fig. 4.1. FOP label granularity. At the top are Healthy Start Rating (4 chunks) and Warning Labels (1 chunk). In botton, Multiple Traffic Light (4 chunks) and Nutri-Score (5 chunks) were used. In the middle, KH (1 level).

(KH) and Warning Labels (WL); and summary indicators such as the Nutri-Score (NS) (see Figure Fig. 4.1). Nutrient-specific systems and summary labels integrate numerical information and visually accessible elements, in contrast with warning labels, which succinctly convey a product's overall health without delving into comprehensive details [15, 24, 25]. FOP labels are designed to condense complex nutrient declarations into intuitive summaries that help consumers identify healthier products at a glance. While prior work has examined colour, shape, and evaluative framing, far less attention has been given to granularity—the degree of detail or the number of informational "chunks" contained in a label [7, 8, 66]. Conceptually, granularity captures how finely a scheme partitions the healthfulness continuum; for example, Nutri-Score and Multiple Traffic Lights categorize products into five ordered levels, whereas warning symbols typically offer only one or two evaluative cues. Early evidence indicates that more finely grained schemes outperform binary warnings in nudging choices toward healthier options because additional strata convey nuance that aids nutritional understanding [15]. However, the cognitive cost of such detail has rarely been considered [107, 108, 109, 110]. Cognitive-load theory posits that working memory can handle only a limited number of chunks at any moment; exceeding this capacity impairs comprehension and recall [111]. Classical

estimates place this limit at approximately seven items [112], although conventional resource models emphasize flexible allocation rather than a fixed span [113]. Consequently, a label that is optimal for a high-capacity consumer may overwhelm a shopper with lower capacity. Despite this theoretical importance, the literature offers little guidance on exactly how many chunks maximize usability across heterogeneous audiences, leaving regulators to choose granularity largely on pragmatic or political grounds. The present study addresses this gap by testing whether the effectiveness of FOP labels depends on, and can be optimized for, consumers' working-memory capacity. Granularity was manipulated via two calorie labels, a coarse, four-chunk format and a detailed, eight-chunk format, which were chosen to bracket the five-chunk MTL standard familiar to UK shoppers. The participants' working memory was indexed with a three-stage n-back test, permitting a direct examination of the label and cognitive-capacity interaction. Formally, we hypothesized a moderation effect whereby the direction and magnitude of the granularity impact would vary as a positive function of n-back performance. We anticipate a cognitive stratification pattern in which coarse labels promote healthier choices among individuals with lower working memory due to their reduced cognitive demands. Conversely, detailed labels should benefit those with higher capacity by offering more comprehensive diagnostic information. By integrating granularity with cognitive-load theory, our study contributes both conceptually and practically. Conceptually, it reframes label design as a problem of information-chunk optimization under resource constraints. Practically, the findings can inform the European Union's ongoing deliberations on a harmonized FOP scheme by specifying when additional detail helps and when it hinders consumer decision-making.

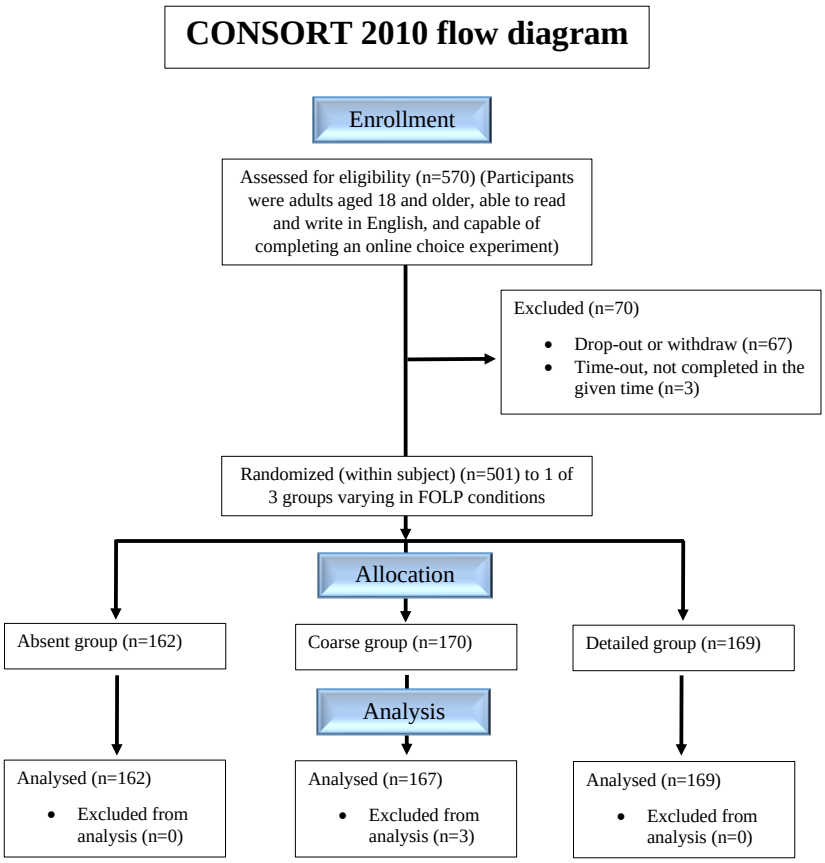

Fig. 4.2. Consort flow diagram reporting recruitment and randomization

4.2 Methods

4.2.1 Study design

We conducted a randomized controlled trial to investigate the influence of FOP labels on individual breakfast cereal choices in an online choice setting. The participants were randomly assigned to one of three groups: a coarse labelling group, in which cereals displayed less granular labelling; a detailed labelling group, in which products conveyed more granular labelling; and a control group, in which products did not display labels. The allocation was concealed from the participants. Breakfast cereals were selected for this research because of their widespread consumption and the established variability in their nutritional content [13]. The online experimental design allowed for the controlled presentation of nutritional information, mimicking a real-world grocery shopping scenario. Prolific Limited (Ltd.) recruited participants, providing a national representative probabilistic sample in Great Britain. As an incentive, the subjects received a £4.5 upon completion. They completed the choice experiment on computer devices; smartphones and tablets were not allowed. The experiment was designed on the Qualtrics online platform via several randomization features.

4.2.2 Outcome measures

The primary outcomes are the average calorie counts of selected products and the probability of selecting lower-calorie products. These measures directly reflect the objective understanding and impact of FOP labels on consumer choices. Additionally, subjective understanding of labelling, assessed through a postexperiment questionnaire, is a secondary outcome.

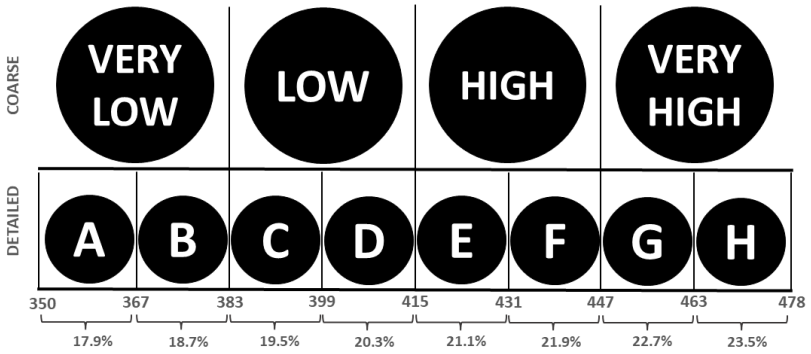

Fig. 4.3. FOP labelling granularity in the choice task, presenting calories per 100 grams and the percentage of average daily reference intake (%RI). The "Coarse" condition uses four categories (very low, low, high, very high), while the "Detailed" condition uses an eight-category letter scale (A–H), with each letter corresponding to a specific calorie range (e.g., A: 350–366 kcal; B: 367–382 kcal).

4.2.3 FOP label granularity

We designed a coarse and detailed label mirroring the structure employed in prior research [114, 17] while adapting the caloric ranges to reflect the breakfast cereals available in this choice experiment (see Figure Fig. 4.3). For example, A represented approximately 358 kcal, and H represented approximately 471 kcal. To contextualize the calorie information, participants received instructions explaining that a 100 g serving of cereal labelled high would provide, on average, 22% of the recommended daily caloric intake on the basis of a 2,000 kcal diet, whereas a serving labelled A would provide 17.9% of the recommended daily caloric intake. Both labels featured a black circular shape with white text and designs created with Adobe Photoshop software.

4.2.4 Procedures

We used a similar methodology to that used in previous studies [34, 38, 17]. The study utilized a choice experiment comprising 16 trials, where participants selected breakfast cereals from an online grocery shopping setting. Table Table 4.1 presents the eight products included in the experiment, which span a caloric range of 360–470 kcal per 100 g, with the nutritional content verified via manufacturer websites. In this study, low-calorie products are defined as those containing 383 kcal or fewer per 100 g on the basis of United Kingdom (UK) regulations [115]. We established this threshold by taking into account the typical serving sizes for breakfast cereals and with the aim of encompassing products at the lower end of the calorie range within the sample. Cereal brands 1 and 2 from Table Table 4.1 met this criterion and were accordingly classified as lower-calorie alternatives. We employed a restricted factorial design to present participants with sets of four cereal brands in each trial. The study mitigated the potential confounding effect of

Table 4.1. Calorie count for breakfast cereals included in the choice experiment ($n = 8$)

Cereal brand	Calorie count (kcal per 100 g)	Coarse	Detailed
None option	0	-	-
1	360	Very low	A
2	374	Very low	B
3	392	Low	C
4	398	Low	D
5	423	High	E
6	431	High	F
7	453	Very high	G
8	470	Very high	H

product familiarity by employing authentic product images captured by the researchers and eliciting participants' self-reported familiarity with the items (see Figure Fig. 4.4). This approach aims to increase the ecological validity of the online shopping environment [116].

Fig. 4.4. Manipulation of the product front package. A calorie label has been added to the original image.

4.2.5 N-back test

We assessed working memory capacity via a computerized n-back test [117, 118, 119, 120]. The participants were presented with a sequence of letters (A, B, C, D, F, H, K, L, M, N, O, P, Q, R, X) displayed individually in the center of the screen with a font size of 30 pt for 1 second for each interval between stimuli. The test utilized a JavaScript program to generate the randomized presentation order of the stimuli comprising three blocks: 1-back, 2-back, and 3-back. The 1-back block is the simplest level, where the stimulus must be compared to the stimulus that immediately preceded it. Examples were provided to the participants prior to commencing the test. The scoring was based on the performance of the participant, specifically concerning correct identifications (hits) when the stimulus was not presented and when the participant responded positively (false alarms). Each block consisted of 30 trials, with a target of 30 hits each, and did not provide any feedback to participants but provided some examples. The performance was assessed via the d-prime (d') measure, which was calculated as $d' = Z(hits) - Z(false\,alarms)$, where Z represents the inverse of the cumulative normal distribution function applied to the hit rate and false alarm rate, respectively [117]. All the participants began with the 1-back block. To progress to the following blocks, they were required to make fewer than 25 errors in the preceding n-back level. Errors were defined as the sum of false alarms and missed responses to target stimuli (miss). This approach helps mitigate the potential confounding

effects of boredom or disengagement on performance in the more demanding stages of the test, a factor particularly relevant in online research environments [121].

4.2.6 Subjective understanding and behavioral questionnaire

We included a brief questionnaire to assess participants' subjective understanding of the FOP labels presented in the choice experiment. The participants reported whether they had sufficient information to make informed choices and whether they recalled seeing the labelling. This approach is similar to that used by [38], where the perceptions of the labels were also assessed. Furthermore, the participants completed a behavioral questionnaire to evaluate their food-shopping behaviors, dietary knowledge, and demographic characteristics.

4.2.7 Statistical analysis

This study used multilevel logistic regression to evaluate the probability that participants would choose lower-calorie options and multilevel linear regression to estimate the calorie counts among the experimental groups. The model specification is as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \text{FOP}_{ij} + \beta_2(\text{FOP}_{ij} * \text{n-back}_{ij}) + \beta_3 \text{Controls}_{ij} + u_j \sim N(0, \sigma_u^2) \tag{4.1}$$

with p_{ij} being the probability that the i trial within the j participant results in the choice of a lower-calorie product. where β_0 represents the intercept (logarithmic odds of the outcome when all predictors equal zero); β_1 captures the effect of different FOP; β_2 estimates the interaction effect between FOP and the n-back test; and β_3 represents control variables that include the choice sets (Trials) and cereal preferences. The term $u_j \sim N(0, \sigma_u^2)$ denotes the participant-specific random intercept, which accounts for the correlation structure induced by the 16 repeated choice trials nested within each individual. This random effect captures unobserved heterogeneity between participants and controls for the

nonindependence of observations, thereby providing more accurate standard errors for the fixed effects. The variance parameter σ_u^2 quantifies the variability between participants in the baseline propensity to choose lower-calorie options after accounting for the fixed effects in the model. Similarly, multilevel linear regression models were used to compare the calorie counts of the selected products under experimental conditions. The outcome variable Y_{ij} in these models represented the average calorie count of the products chosen by participant j in trial i, including an error term e_{ij}. In all regressions, standard errors were clustered at the participant level. The data collection process excluded participants if they self-reported not buying or eating cereal in the past twelve months to ensure that the lack of familiarity with the products did not affect the findings. The analyses were not blinded but were carried out by a statistician-in-training who had participated in the preparation and conduct of the experiment. All analyses were performed via R statistical software [91], and we reported a significance level of 5%.

4.3 Results

4.3.1 Participant characteristics

The participant flowchart for recruitment and randomization is presented in Figure Fig. 4.2. Data were collected from 29th August–10th September 2024 from participants aged 18 years and older who were able to complete an online experiment. Prolific Ltd. invited 570 panel members to participate, and the complete data of 498 participants were considered in the analysis. This represents 87.3% of the recruited sample. Of the original participants, 67 withdrew from the study, and data from 3 participants were excluded from the analysis because they exceeded the maximum allotted time for providing responses. Individuals were randomly divided into three different experimental conditions, and the sample sizes were similar across groups: absent (n = 162), coarse (n = 170) and detailed (n = 169). Three participants were excluded from the detailed group after randomization because of

incomplete data. Individuals took 22 minutes to complete the randomized control trial on average, which included the choice task, n-back test, and posterior behavior questionnaires. Table Table 4.3 presents the household composition and food purchasing behaviors of the

Table 4.2. Participant characteristics by experimental group ($n = 498$)

Variable	Alternative	All	Absent n (%)	Coarse n (%)	Detailed n (%)	χ^2 (p)
Participants		498	162 (32)	167 (34)	169 (34)	
Sex	Female	257 (52)	86 (54)	91 (55)	80 (47)	0.39
Sex	Male	240 (48)	75 (46)	75 (45)	89 (53)	
Sex	Unreported	1 (0)	1 (0)			
Age	18-34	140 (28)	44 (27)	46 (28)	50 (29)	0.94
Age	35-54	165 (33)	55 (34)	58 (35)	52 (31)	
Age	55-65+	193 (39)	63 (39)	63 (37)	67 (40)	
Ethnicity	White	417 (84)	138 (85)	139 (83)	140 (83)	0.89
Ethnicity	Mixed	75 (15)	23 (14)	26 (16)	26 (15)	
Ethnicity	Unreported	6 (1)	1 (1)	2 (1)	3 (2)	
Education	Higher	199 (40)	59 (36)	66 (40)	74 (44)	0.57
Education	Lower	297 (60)	102 (63)	101 (60)	94 (56)	
Education	Unreported	2 (0)	1 (1)	0 (0)	1 (1)	
Income[1]	Higher	267 (54)	93 (57)	96 (58)	78 (46)	0.13
Income	Lower	195 (39)	60 (37)	57 (34)	78 (46)	
Income	Unreported	36 (7)	9 (6)	14 (8)	13 (8)	
BMI [2]	Obesity	130 (26)	45 (28)	38 (23)	47(28)	0.95
BMI	Overweight	159 (32)	52 (32)	55 (33)	52 (31)	
BMI	Normal weight	191 (38)	59 (36)	68 (41)	64(38)	
BMI	Underweight	18 (4)	6 (4)	6 (4)	6 (4)	

[1] Higher income: £30,001 - Above £40,000; Lower income: Below £10,000 - £30,000. [2] BMI (body mass index) was estimated with self-reported height and weight.

experimental groups. Approximately one-third of the participants had children in the household (32%), and the majority had food-shopping responsibilities (always = 51%, most of the time = 27We evaluated participants' preferences for the cereal brands in the choice experiment by asking them to complete a ranking task, where they ordered the products on a Likert scale (1= most preferred to 8= least preferred). Table Table 4.4 shows the mean rank and standard deviation (SD) for each cereal brand; lower mean ratings indicate greater preference. Cereal brand 2 (mean: 3.60; SD: 2.60) was the most preferred average, whereas brand 7 (mean: 5.91; SD: 2.33) was the least preferred. The ranking metrics for product preferences were not similar across the experimental groups. The preference ranking for cereal brand 4 differed significantly across the experimental

Table 4.3. Household composition and purchasing behaviors by experimental group ($n = 498$)

Variable	Alternative	All	Absent n (%)	Coarse n (%)	Detailed n (%)	χ^2 (p)
Participants		498	162 (32)	167 (34)	169 (34)	
Children at home	No	334 (67)	103 (64)	108 (65)	123 (73)	0.39
Children at home	Yes	157 (32)	56 (35)	57 (34)	44 (26)	
Children at home	Unreported	7 (1)	3 (1)	2 (1)	2 (1)	
Losing weight	No	224 (45)	70 (43)	75 (45)	79 (47)	0.47
Losing weight	Yes	268 (54)	91 (56)	91 (54)	86 (51)	
Losing weight	Unreported	6 (1)	1 (1)	1 (1)	4 (2)	
Responsibility	Always	255 (51)	83 (51)	87 (52)	85 (50)	0.99
Responsibility	Most of the time	135 (27)	45 (28)	45 (27)	45 (27)	
Responsibility	Half of the time	60 (12)	18 (11)	20 (12)	22 (13)	
Responsibility	Sometimes	43 (9)	15 (09)	13 (07)	15 (08)	
Responsibility	Never	5 (1)	1 (1)	2 (1)	2 (1)	
Check ingredients	Always	58 (12)	23 (14)	20 (12)	15 (09)	0.07
Check ingredients	Most of the time	177 (36)	58 (36)	70 (42)	49 (29)	
Check ingredients	Half of the time	111 (22)	31 (19)	36 (22)	44 (26)	
Check ingredients	Sometimes	132 (27)	43 (27)	33 (20)	56 (33)	
Check ingredients	Never	20 (4)	7 (4)	8 (8)	5 (3)	
Familiarity	Very familiar	174 (35)	52 (32)	70 (41)	52 (31)	0.20
Familiarity	Somewhat familiar	272 (55)	90 (56)	82 (50)	100 (60)	
Familiarity	Neither	20 (4)	6 (04)	5 (03)	9 (05)	
Familiarity	Unfamiliar	32 (6)	14 (8)	10 (6)	8 (4)	

conditions according to an analysis of variance (ANOVA $p < 0.05$). Analyses revealed that participants in the coarse group (mean: 3.75; SD: 1.95) ranked this product lower than those in the absent condition (mean: 4.31; SD: 1.94). Similarly, cereal 6 showed differences in preference rankings across conditions (ANOVA $p < 0.05$). The participants in the coarse condition (mean: 4.26; SD: 1.86) showed a significantly greater preference for cereal 6 than did those in the absent condition (mean: 3.61; SD: 1.88). For cereal 8, the

Table 4.4. Product preferences by experimental group, means and SDs ($n = 498$)

Cereal brand	Calorie count	Absent n (%)	Coarse n (%)	Detailed n (%)	Anova (p)
Participants		162 (0.32)	167 (0.34)	169 (0.34)	
1	360	4.31 (2.37)	4.07 (2.42)	3.98 (2.36)	0.43
2	374	3.49 (2.49)	3.39 (2.56)	3.91 (2.74)	0.15
3	392	4.98 (1.83)	4.55 (2.04)	4.66 (1.93)	0.12
4	398	4.31 (1.94)	3.75 (1.95)	3.93 (2.00)	0.03
5	423	5.47 (2.03)	5.32 (1.97)	5.31 (2.17)	0.74
6	431	3.61 (1.88)	4.26 (1.86)	3.96 (1.79)	0.00
7	453	5.94 (2.37)	6.04 (2.24)	5.76 (2.37)	0.54
8	470	3.88 (2.13)	4.60 (2.07)	4.49 (2.11)	0.00

detailed condition also significantly affected the preference rankings (ANOVA $p < 0.05$). The participants in the coarse (mean: 4.60; SD: 2.07) and detailed conditions (mean: 4.49; SD: 2.11) showed significantly greater preferences than did those in the absent condition (mean: 3.88; SD: 2.13).

4.3.2 N-back performance across experimental groups

Table Table 4.5 presents the distributions of participants' *d'* scores across the experimental groups for each level of the n-back block. To account for potential boundary issues where *d'* scores might be undefined owing to hit or false-alarm rates of 0 or 1, we applied the correction method described by Stanislaw and Todorov [122]. Specifically, any score of 0 was replaced with 0.5/n, and any score of 1 was replaced with $(n - 0.5)/n$, where n represents the number of target and noise trials. In this study, the target was equal to 30, as each block contained 30 trials, and the noise was established at 48, the highest observed false alarm rate across all participants. The sample sizes for the 1-back, 2-back, and 3-back conditions were 498, 464, and 452, respectively. This represents a reduction of 7.33% and 10.18% from the initial sample size, likely due to participant data exclusion on the basis of predefined criteria. As hypothesized, n-back test performance declined as task difficulty increased. The 3-back condition presented the lowest mean *d'* scores across all treatment groups (absent: 2.29 (SD 0.87), coarse: 2.29 (SD 0.82), and detailed: 2.16 (SD 0.90)). While the mean *d'* scores and standard deviations were relatively consistent across groups, no differences were observed in scores between groups (ANOVA $p > 0.05$). The total duration of the n-back test averaged approximately 11 minutes (SD = 3.25 minutes) across all groups.

4.3.3 Treatment effect on calorie counts

The results from multilevel linear regression analyses examining the effects of FOP labels on calorie counts are presented in Table Table 4.6. All the models were adjusted for choice

Table 4.5. *d'* scores by experimental group, *n* (Mean, SD)

Variable	1-back (n=498)	2-back (n=464)	3-back (n=452)
Absent	162 (3.37, 1.01)	157 (3.24, 1.00)	150 (2.29, 0.87)
Coarse	167 (3.13, 1.26)	150 (3.24, 0.91)	148 (2.29, 0.82)
Detailed	169 (3.21, 1.17)	157 (3.14, 0.93)	154 (2.16, 0.90)
Anova (p)	0.16	0.59	0.30

sets (Trials) and product preferences. The baseline model (Model 1) indicated negative but not significant direct effects of either Coarse (-9.14, standard error [SE] = 7.20) or Detailed labels (-3.46, SE = 7.18) on calorie counts. These effects remained stable after controlling for choice sets (Model 2) and product preferences (Model 3). However, the introduction of n-back performance interactions revealed notable heterogeneous effects. Particularly compelling were the significant negative interactions between labelling and 2-back performance (Model 5): both Coarse*2-back (-14.51, SE = 6.90, $p < 0.05$) and Detailed*2-back (-17.35, SE = 6.75, $p < 0.05$) interactions indicate that as participants' performance on the n-back test improves (reflected by a unit increase in d'), the average caloric count decreases. This pattern persisted in the 3-back block (Model 6), where detailed labels had a stronger moderating effect (-18.05, SE = 7.42, $p < 0.05$) than coarse labels did (-13.52, SE = 7.83, $p < 0.1$). The 1-back block showed similar but less pronounced interaction patterns. In Model 6 and Figure Fig. 4.5 A), the Absent group (intercept) had an average calorie count of 361.93 kcal, the Coarse*3-back group averaged 348.41 kcal (3.88% decrease versus Absent), and the Detailed*3-back group averaged 343.88 kcal (5.24% decrease versus Absent). Figure Fig. 4.5 B) displays nonparallel trend lines for the different labelling conditions, indicating that the impact of the labelling on the calorie count varies across the *d'* values. Table Table 4.7 presents Hedges-adjusted Cohen's *d* values alongside model-based means and their 95% confidence intervals. Across the working-memory spectrum *d'*, most Cohen's *d* values fluctuate between –0.30 and +0.30, with nearly every confidence band intersecting 0, indicating that the observed differences are not only quantitatively small but also statistically imprecise. A notable exception is the Detailed-versus-Absent contrast at the highest 3-back level, where the

Table 4.6. Multilevel linear regression results: effects of FOP labelling on calorie count by experimental conditions, including n-back test performance levels as an interaction term (adjusted for product preferences and choice sets (Trials)).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	372.70***	388.37***	343.70***	319.15***	341.83***	361.93***
	(5.13)	(5.47)	(43.49)	(45.20)	(45.35)	(44.64)
Coarse	−9.14	−9.14	−2.55	31.28	43.70*	26.06
	(7.20)	(7.20)	(6.74)	(20.57)	(23.28)	(19.13)
Detailed	−3.46	−3.46	−2.73	46.51**	48.53**	30.66*
	(7.18)	(7.18)	(6.66)	(21.14)	(22.55)	(17.74)
1-back				6.40		
				(4.66)		
Coarse*1-back				−10.25*		
				(5.92)		
Detailed*1-back				−15.00**		
				(6.08)		
2-back					5.30	
					(4.62)	
Coarse*2-back					−14.51**	
					(6.90)	
Detailed*2-back					−17.35**	
					(6.75)	
3-back						1.03
						(5.38)
Coarse*3-back						−13.52*
						(7.83)
Detailed*3-back						−18.05**
						(7.42)
Choice sets		*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Preferences			*Yes*	*Yes*	*Yes*	*Yes*
AIC	95763.53	95698.97	95663.85	95583.60	89133.72	86821.42
Num. obs.	7968	7968	7968	7968	7424	7232
Num. groups	498	498	498	498	464	452

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$

effect reaches $d = -0.28$ (95% CI [0.50, 0.06]); this approximates the lower threshold of a "small" effect by conventional standards. However, the upper limit remains considerably below the threshold for a moderate effect, suggesting only a modest practical benefit.

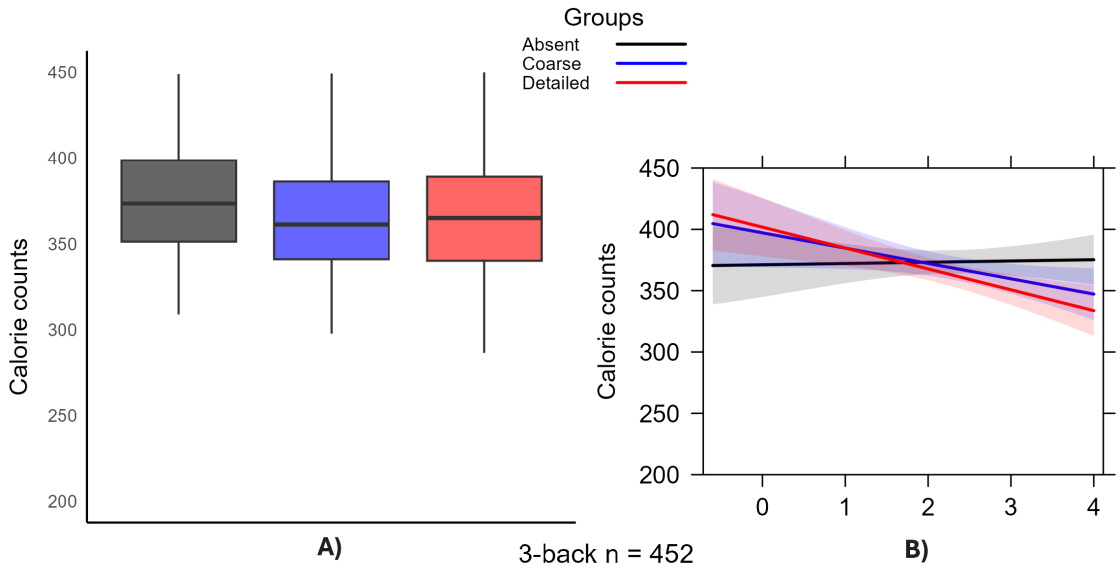

Fig. 4.5. A) Effects of FOP labelling on calorie counts by experimental conditions, including 3-back test performance levels as an interaction term. B) Plot showing the interaction effects between d' as 3-back performance and different labelling conditions on caloric count.

Index	*d'*	Group	Mean (95% CI)	Cohen's *d* (95% CI)	
				vs. Absent	vs. Coarse
2-back	2.26	Absent	367.2 (354.6, 379.8)	â€“	+0.12 (â€“0.10, 0.34)
		Coarse	378.1 (364.5, 391.7)	â€“0.12 (â€“0.34, 0.10)	â€“
		Detailed	376.5 (364.0, 389.0)	â€“0.10 (â€“0.32, 0.12)	+0.02 (â€“0.20, 0.24)
2-back	3.21	Absent	372.3 (363.2, 381.3)	â€“	â€“0.03 (â€“0.25, 0.19)
		Coarse	369.4 (360.1, 378.6)	+0.03 (â€“0.19, 0.25)	â€“
		Detailed	365.1 (356.1, 374.1)	+0.08 (â€“0.14, 0.30)	+0.05 (â€“0.17, 0.27)
2-back	4.15	Absent	377.2 (364.9, 389.6)	â€“	â€“0.18 (â€“0.40, 0.04)
		Coarse	360.7 (347.7, 373.7)	+0.18 (â€“0.04, 0.40)	â€“
		Detailed	353.8 (340.5, 367.0)	+0.25 (0.03, 0.47)	+0.07 (â€“0.15, 0.29)
3-back	1.38	Absent	372.5 (359.4, 385.7)	â€“	â€“0.08 (â€“0.30, 0.14)
		Coarse	379.9 (366.1, 393.8)	+0.08 (â€“0.14, 0.30)	â€“
		Detailed	378.3 (366.4, 390.2)	+0.06 (â€“0.16, 0.28)	â€“0.02 (â€“0.24, 0.20)
3-back	2.25	Absent	373.4 (364.3, 382.6)	â€“	+0.05 (â€“0.17, 0.27)
		Coarse	369.1 (359.9, 378.3)	â€“0.05 (â€“0.27, 0.17)	â€“
		Detailed	363.5 (354.5, 372.5)	â€“0.11 (â€“0.33, 0.11)	â€“0.06 (â€“0.28, 0.16)
3-back	3.11	Absent	374.3 (361.6, 387.0)	â€“	+0.17 (â^'0.05, 0.39)
		Coarse	358.3 (345.4, 371.3)	â€“0.17 (â^'0.39, 0.05)	â€“
		Detailed	348.8 (335.8, 361.9)	â€“0.28 (â^'0.50, â^'0.06)	â€“0.11 (â^'0.33, 0.11)

Table 4.7. Predicted cereal–selection position and pairwise Cohen's *d* by *d'* level.

4.3.4 Treatment effect on the probability of choosing lower-calorie products

Table Table 4.8 presents the multilevel log-binomial regression results, including interactions between the labelling groups and the 3-back block adjusted for controls. The analysis spans the full range of available calorie options, from the lowest-calorie (360 and 374 kcal) to the highest-calorie products (453 and 470 kcal). For the lowest-calorie options, contrasting patterns emerge. At 360 kcal, the Coarse*3-back interaction is negative and significant (-0.80, SE = 0.36, $p < 0.05$), indicating that participants with higher working memory capacity, reflected by a unit increase in *d'*, were less likely to select this low-calorie option when presented with coarse labels. The Detailed*3-back interaction has no significant effect (0.11, SE = 0.33). In the moderate-calorie range (392 kcal), both interaction terms are positive and significant (Coarse*3-back 0.69, SE = 0.41, $p < 0.1$; Detailed*3-back 0.90, SE = 0.39, $p < 0.05$), suggesting that participants with higher *d'* values were more likely to select this moderate-calorie option when presented with either type of labelling. For the 398-calorie option, neither interaction term reached statistical significance (Coarse*3-back 0.01, SE = 0.61; Detailed*3-back 0.06, SE = 0.58). For higher-calorie options (423–470 kcal), none of the interaction terms reached statistical significance, although there was a trend toward negative coefficients for Detailed*3-back interactions, particularly for the 423-calorie (-0.48, SE = 0.32) and 431-calorie (-0.51, SE = 0.38) options. See Appendix B for the results of the 2-back and 1-back blocks, which showed similar directional patterns to the 3-back findings for lower-calorie products (360 and 374 kcal), but with notably weaker effects. This suggests that the influence of FOP labels on lower-calorie product selection was most pronounced under higher cognitive load conditions.

Table 4.8. Multilevel log-binomial regression results—effects of FOP labelling on the probability of choosing lower-calorie cereal brands by experimental groups with the 3-back level as the interaction term (adjusted for product preferences and choice sets (Trials)).

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Cereal brands	None	360	374	392	398	423	431	453	470
(Intercept)	−7.32**	−4.64***	6.39***	−1.38*	−6.12***	−3.34***	−2.14***	−6.42***	−8.61***
	(3.05)	(0.66)	(0.53)	(0.76)	(1.20)	(0.65)	(0.74)	(1.41)	(1.44)
Coarse	−1.92	1.77**	0.83	−0.41	0.91	−0.07	−2.59**	−0.40	−0.47
	(1.29)	(0.85)	(0.65)	(0.98)	(1.47)	(0.87)	(1.05)	(1.65)	(1.87)
Detailed	−1.39	−0.13	0.39	−0.93	0.61	0.56	−0.07	0.40	−0.14
	(1.21)	(0.78)	(0.61)	(0.92)	(1.39)	(0.77)	(0.90)	(1.51)	(1.69)
3-back	0.04	0.02	0.09	−0.44	−0.43	0.54**	0.02	−0.13	−0.03
	(0.36)	(0.25)	(0.19)	(0.29)	(0.42)	(0.23)	(0.26)	(0.46)	(0.51)
Coarse*3-back	0.89*	−0.80**	−0.25	0.69*	0.01	−0.26	0.46	0.14	0.00
	(0.52)	(0.36)	(0.27)	(0.41)	(0.61)	(0.35)	(0.42)	(0.68)	(0.77)
Detailed*3-back	0.78	0.11	−0.12	0.90**	0.06	−0.48	−0.51	−0.05	−0.13
	(0.50)	(0.33)	(0.25)	(0.39)	(0.58)	(0.32)	(0.38)	(0.63)	(0.71)
AIC	2926.66	2524.06	3012.73	4083.45	2897.13	4408.81	3615.88	2313.45	2213.25
Num. obs.	7232	7232	7232	7232	7232	7232	7232	7232	7232
Num. groups	452	452	452	452	452	452	452	452	452

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 4.9. Cohen's d for interaction terms (multilevel logistic models)

Cereal brands	Coarse*3-back	Detailed*3-back
None	0.49 [−0.07, 1.05]	0.43 [−0.11, 0.97]
360	−0.44 [−0.83, −0.05]	0.06 [−0.30, 0.42]
374	−0.14 [−0.43, 0.15]	−0.07 [−0.34, 0.20]
392	0.38 [−0.06, 0.82]	0.49 [0.08, 0.92]
398	0.01 [−0.65, 0.67]	0.03 [−0.59, 0.66]
423	−0.15 [−0.52, 0.24]	−0.27 [−0.61, 0.08]
431	0.25 [−0.20, 0.71]	−0.28 [−0.69, 0.13]
453	0.08 [−0.66, 0.81]	−0.03 [−0.71, 0.65]
470	0.00 [−0.83, 0.83]	−0.07 [−0.84, 0.70]

4.3.5 Subjective Understanding

Table Table 4.10 shows the results of the subjective FOP understanding questions by experimental groups. The participants in the coarse labelling condition consistently demonstrated the highest FOP recall rate, reaching 48.0

Table 4.10. Subjective FOP understanding questions by experimental group ($n = 498$)

Variable	Alternative	All	Absent	Coarse	Detailed	χ^2 (p)
Participants		498	162 (32)	167 (34)	169 (34)	
Reported seeing FOP	Always	147 (44)		80 (48)	67 (40)	0.22
Reported seeing FOP	Most of the time	104 (31)		51 (36)	53 (31)	
Reported seeing FOP	About half the time	32 (10)		10 (6)	22 (13)	
Reported seeing FOP	Sometimes	39 (11)		19 (11)	20 (12)	
Reported seeing FOP	Never	14 (4)		7 (4)	7 (4)	
Having information	Overload or sufficient	244 (49)	54 (33)	99 (60)	91 (54)	0.00
Having information	Neither	66 (13)	26 (16)	19 (11)	21 (12)	
Having information	Somewhat insufficient	133 (27)	51 (31)	36 (22)	46 (27)	
Having information	Insufficient	55 (11)	31 (19)	13 (8)	11 (7)	

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

4.3.6 Robustness checks

The multilevel linear regression analysis on calorie count investigated six specific groups Ã—n-back interactions (see Appendix D, Table 1). Prior to adjusting for multiple comparisons, four of these six interactions reached nominal significance. However, after applying Holm's familywise correction criterion at $\alpha = .05$, none of the variables remained significant. Nevertheless, three interactions involving the detail-label condition retained significance under the less stringent Benjamini–Hochberg procedure. In the multilevel log-binomial models, we examined 16 interaction contrasts (see Appendix D, Table 2). None of the 16 contrasts survived Holm's correction, and only the Detailed*3-back interaction at 392 kcal remained significant at the 5% false discovery rate (FDR) threshold. This attenuation likely stems from the statistical penalty associated with evaluating numerous contrasts on a limited sample rather than a complete absence of genuine effects. Therefore, the corrected results should be interpreted as conservative estimates, with effect size estimates and their confidence intervals providing a clearer indication of practical significance. To assess the robustness of our findings to downsampling across n-back blocks, we performed a sensitivity analysis via a matched sample approach. Specifically, we rerun all multilevel regression models using only participants (N = 452) who completed the three n back blocks, thus holding the sample size constant across working memory conditions. The

results from these models were highly consistent with those from the full-sample analyses (see Appendix A). The direction and magnitude of the n-back coefficients remained stable, indicating that the relationship between working memory and FOP effects is not driven by differential dropout. However, the confidence intervals were wider in the full sample, and only the interaction between the detailed label condition and 3-back performance remained statistically significant in the matched sample. This pattern is expected given the reduced sample size and associated loss of statistical power. Furthermore, the completion rate for the sequence of n-back tasks was notably high and was largely independent of the demographic characteristics of the participants (see Appendix E). Chi-square analyses comparing individuals who completed the entire sequence with those who discontinued after the 1 or 2 back blocks revealed no significant differences across the variables examined. Although ethnicity demonstrated a statistically significant association, the effect size was small and diminished to nonsignificance upon collapsing sparse categories, suggesting that the result was driven by low cell counts rather than a systematic pattern of dropout. Collectively, these results suggest that attrition across the n-back blocks was minimal and effectively random with respect to the demographics measured, thus reducing concerns that bias from selective dropout could influence working memory analyses.

4.4 Discussion

This randomized controlled trial offers preliminary evidence suggesting that the effectiveness of FOP labels is partially dependent on and can be optimized for consumer working memory capacity. Across the choice experiment, enhanced n-back performance was consistently correlated with reduced calorie selection under both coarse and detailed label formats, with the most pronounced decrease observed during the demanding 3-back block. Calorie-count models indicated that each one-unit increase in 3-back *d'* reduced purchase calories by approximately 14 kcal with coarse labels and 18 kcal with detailed labels. The Detailed-versus-Absent contrast at the highest capacity level yielded an effect

size of –0.28, which is small but practically significant when it is extrapolated over multiple shopping instances. Multilevel log-binomial regression analyses further indicated that individuals with greater capacity were less likely to select the lowest-calorie option when viewing coarse labels but were more inclined to choose a moderate-calorie product when viewing detailed labels. While these interactions are context-specific and generally small, they suggest a consistent pattern rather than chance. These findings partially validate the role of cognitive stratification in label effectiveness. They indicate that consumers may employ different processing strategies on the basis of their working memory capacity and the level of detail provided by the labels, as recent work has validated. For example, a comprehensive post hoc analysis of an international cross-sectional survey with a sample of 3,680 adults across 18 nations revealed parallel findings: less than 50% noticed the detailed nutrient-warning FOP during simulated choices, and of those, only one-third could identify the least-healthy option [123]. This suggests that the diagnostic advantages of detailed warning schemes may diminish when label salience or cognitive resources are limited. Similarly, eye tracking in an augmented-reality grocery setting shows that shoppers devote 20–30% fewer fixations to the MTL labels rather than simple "healthy choice" icons, which we interpret as a consequence of the cognitive demand [108]. Furthermore, a recent meta-analysis focusing on individuals aged 16 to 35 years revealed that granular-indicator systems yield the most consistent reductions in calorie selection, followed by color-coded and positive-logo designs; conversely, purely numerical formats have minimal impacts on healthier purchasing [110]. These collective findings emphasize that detailed labels enhance dietary outcomes only when the target demographic possesses both the opportunity and the cognitive capacity to process the information [108, 109]. A total reduction of 4.56% in the average calorie count was observed when the FOP labels were combined with those of the control group, considering 3-back as a moderator. This variation represents a decrease of 3.88% in the coarse group and 5.24% in the detailed group, where 4–5% represents approximately 100–110 calories of a typical 2.000 calorie diet. This reduction is comparable to skipping a 330 ml sugar-sweetened beverage from daily intake, which typically contains approximately 35–40 grams of free sugars. Population modelling studies

suggest that reducing daily free sugar intake by 30 g can lead to an average body weight reduction of 1.2–1.6 kg over one year and decrease the incidence of type 2 diabetes by 2–3% [124, 125]. An observational study estimated a similar calorie reduction of 6.5% in calories from purchased cereals after the introduction of compulsory warning labels in Chile [13]. Likewise, research conducted in the United States of America indicated a 3% reduction in the average calorie amount under comparable FOP label conditions during the product selection task [17]. In contrast, the results of this study revealed that Coarse was associated with greater calorie reduction than was the control. The findings of the present study diverge from ours, which can be attributed to variations in measurement approaches. For moderate-calorie products, the positive and statistically significant interaction terms indicate that participants with higher 3-back *d'* scores were more inclined to choose this mid-range product when either label was present. However, at a slightly elevated caloric level of 398 kcal, the interactions were not significant. Furthermore, for the higher-calorie range, none of the interaction terms reached statistical significance, although the Detailed*3-back coefficients consistently exhibited a negative trend. Collectively, these findings imply that FOP labels exert their most pronounced capacity-dependent effect by guiding individuals with high cognitive capacity toward moderate-calorie products rather than discouraging them from lower-calorie choices altogether [8, 66, 15, 126]. These mixed results may indicate that individuals' choices are influenced by their preexisting beliefs about the calorie content of products [9]. Consistent with previous studies, labelling may be most effective for moderately caloric products for which consumers hold inaccurate calorie beliefs, and memory capacity could enhance the effectiveness even for more granular labelling formats [13]. Sensitivity checks indicated that our results are statistically conservative. First, even when a correction for multiple comparisons in the calorie count models was employed, not all the interactions maintained significance, highlighting a modest capacity-dependent advantage of the more granular format. Similarly, within the log-binomial choice models, the detailed*3-back term for the 392 kcal option remained significant at the 5% FDR. While these adjustments reduce statistical power when examining numerous contrasts within a limited sample, the direction

and magnitude of the unrelated effects still indicate small but practical impacts. Second, the effect sizes and directions remained largely consistent, as did the detailed*3-back interaction, upon rerunning all multilevel regressions in a matched subsample. Granular FOP labels can also directly impact consumers' subjective understanding, influencing their perceptions of how well they have comprehended the information from the label. Descriptive analysis revealed that labelling was generally perceived favourably in terms of visibility and provided sufficient information for decision-making. This factor does not appear to have contributed to the comparative effects between coarse and detailed labelling. Notably, 44% of the participants always remembered seeing the FOP label to which they were randomized; recall was highest for the coarse label (48%) and lowest for the detailed label (40Labelling is just one factor among many that the participants considered when selecting cereals. For example, the participants indicated a high degree of familiarity, at a rate of 80%, with all cereal products included in the experiment and, on average, ranked their preferences for these items as 3.59, which reflects a moderately strong preference. This finding indicates a potential link between familiarity, preferences, and the use of nutritional information, which may alleviate the cognitive load associated with processing more comprehensive labels. This finding aligns with previous studies on conjoint experiments, suggesting that FOP labels may influence choices compared with other multiple-product attributes. Previous research has indicated that labelling is one of several factors contributing to parents' choices of children's snack products in the conjoint choice setting, alongside visual information and nutritional claims such as high fibre content [127]. However, further investigations are needed to corroborate these results. These findings enhance the current understanding of FOP labelling policies in three significant ways. First, the data suggest that neither a simplified summary label nor a more granular label provides a universally superior solution. Both label types improved calorie selection, and the hypothesized benefit of reduced granularity was not evident. Second, the evidence indicates that label effectiveness is partially dependent on cognitive resources. Only participants demonstrating higher performance in the 3-back task consistently shifted toward moderate-energy options, suggesting that working-memory capacity

influences the ability to utilize detailed label information. Third, multiple-comparison and matched-sample analyses demonstrate that these capacity-dependent effects, while modest in size, are statistically robust and unlikely to result from selective dropout. Collectively, these results suggest that future European Union labelling policies should move beyond a standardized approach [128, 129]. Policymakers could integrate a single interpretive label with supplementary tools—such as digital overlays or layered information formats—allow consumers with lower numeracy skills or limited working-memory capacity to reduce informational complexity while still providing more detailed data to those who can utilize it [130]. Tailoring label granularity to consumers' cognitive resources could therefore maximize public-health benefits without sacrificing informational completeness.

4.4.1 Limitations

This study has several notable strengths. Our study employed prespecified protocols and analysis plans, which were built upon prior research [18]. We developed a choice task involving a set of cereals rather than individual items, as prior research has mentioned that FOP labels perform more effectively in the context of multiple product options available in a grocery setting. This serves to enhance the external validity of our findings. Additionally, our study focused on examining the isolated effects of label granularity by including no-colour labelling options in the choice task rather than incorporating coloured variants. Despite these advantages, the research design has several limitations. The study incorporated a main effects analysis, employing procedures to control the familywise error rate via the Holm method. This analysis revealed that the contrasts did not reach significance, suggesting that our observed moderated effects should be interpreted cautiously pending further replication. We acknowledge that the online setting, literate and technology-savvy sample, and single food category limit the generalizability of our findings to real-world shopping environments. For example, online settings offer fewer distractions than do busy in-store environments. Online shoppers can easily search and filter products, an option that may not be available in physical stores, and online shoppers

lack the sensory cues that can influence food choices in real life. The experiment did not include back-of-pack information despite its availability in real-world settings where additional data could further inform decision-making. Furthermore, factors such as brand loyalty, pricing, shelf placement, and packaging cues were not considered in our study but could influence consumer behavior in real-world settings. Future research should validate our findings in real-world shopping environments by, for example, conducting experiments in actual grocery stores to observe consumer behavior in a more realistic setting [16]. Other aspects of the study design may have reduced the external validity of the findings. It is possible that the analysis included participants who did not fully complete the working memory test instructions. Excluding these individuals from the sample size did not yield any significant effects; nonetheless, the results should be interpreted as indicative of behavioural trends rather than definitive conclusions. The study employed an n-back test to measure working memory, but alternative approaches could be utilized to relate label granularity effectiveness to cognitive capacity. Finally, despite the researchers' efforts to design images that facilitated accurate selections across all the items, the possibility cannot be ruled out that the detailed labels were unsuitable and challenging for some users, leading to them being overlooked.

4.5 Conclusions

This study provides evidence that FOP calorie labels can significantly influence the energy content of cereal choices. The magnitude and direction of this influence are contingent on both the granularity of the label and the working-memory capacity of consumers. Compared with the absence of labelling, FOP labels reduced average calorie selection by 4.56%, with the detailed format leading to the most substantial reduction and the coarse format resulting in a smaller decrease. Notably, these overall effects concealed a strong capacity-dependent trend: each one-unit increase in 3-back performance corresponded to an additional 14 kcal reduction with coarse labels and 18 kcal with detailed labels. High-

capacity shoppers demonstrated a reduced likelihood of selecting the highest-calorie cereals and an increased preference for moderate-calorie options when presented with detailed labels. Collectively, these results offer partial validation for cognitive stratification theory, suggesting that consumers utilize various decision-making approaches on the basis of their working-memory resources and the informational demands of the packaging. Therefore, adapting label granularity to match consumers' cognitive resources could optimize public-health outcomes without compromising the comprehensiveness of information.

Supplementary information

Appendix A: Table 1. Multilevel linear regression results: effects of FOP labelling on caloric count by experimental conditions, including n-back test performance levels as an interaction term (n = 452). Appendix B: Table 1. Multilevel log-binomial regression results—effects of FOP labels on the probability of choosing lower-calorie cereal brands by experimental groups with the 2-back level as the interaction term. Table 2. Multilevel log-binomial regression results—effects of FOP labels on the probability of choosing lower-calorie cereal brands by experimental groups with the 1-back level as the interaction term. Appendix D: Table 1. Interaction t values with Holm and Benjamini–Hochberg corrections. Table 2. 3-back interaction p values with Holm and Benjamini–Hochberg corrections. Appendix E: Table 3. N-back block completion rates. Table 4 Chi-square comparison of demographics between full and partial completers. Appendix F: Details of the randomized control trial sample and product randomisation.

Ethics approval and consent to participate

This study was conducted in accordance with the guidelines of the Declaration of Helsinki, and all procedures involving human subjects/patients were approved by the University of Manchester Research Ethics Committee (Reference number: 2024–16718–35972). Informed

consent was obtained from all the subjects.

Chapter 5

On the Selection of Covariates for Transportability

Abstract

We generalize experimental nutritional outcomes from a recent breakfast cereal choice trial (N = 498) examining label granularity to the target population represented by the United Kingdom National Diet and Nutrition Survey (N = 10,696). Employing doubly robust Bayesian additive regression tree estimators, we establish that inclusion of a theoretically motivated effect modifierâ€"namely, working memoryâ€"is critical for unbiased estimation of population average treatment effects. Substantively, our analysis uncovers a pronounced backfire effect, whereby detailed labels elicit excess caloric count in the target population, driven by cognitive overload and moral licensing. Incorporating these transported estimates into a 10-year dynamic microsimulation highlights the perils of unmeasured confounding: models based on conventional demographic covariates erroneously forecast benefits from detailed labels, whereas fully cognition-adjusted models reveal net health detriments from consumer-oriented detailed labeling. In contrast, a combined strategy of coarse labels paired with mandatory industry reformulation proves superior and economically feasible. This integrated approach mitigates dietary compensation and behavioral licensing, forecasting the aversion of 439,900 obesity cases, 186,400 quality-adjusted life years gained, and £765.1 million in savings for the National Health Service over 10 years. Our study delivers a practical policy assessment affirming that intervention effectiveness depends on complementary supply-side measures, while illustrating sophisticated transportability methods.

Keywords: Front-of-pack labeling, Nutritional policy, Causal inference, Transportability, Population health, Microsimulation, Obesity prevention.

5.1 Introduction

Front-of-package (FOP) nutritional labelling is a crucial public health strategy designed to reduce the consumption of energy dense foods, which drive the global prevalence of noncommunicable diseases such as obesity [131, 129, 4]. These nutritional labels aim to improve dietary quality by simplifying complex nutritional profiles, thereby enhancing consumer understanding and facilitating healthier purchasing decisions [21, 6]. Consequently, accurately evaluating the cause and effect relationships of varying label designs is essential for optimizing public health interventions [7, 8, 66, 132].

Although substantial evidence supports the efficacy of nutritional labels, this evidence is primarily derived from randomized controlled trials (RCT) conducted in isolated experimental settings [133, 34, 8]. However, population-level assessments that adequately adjust for real-world confounders remain critically scarce. Consequently, the current literature fails to provide precise, population-based projections to guide policymakers in allocating resources to efficiently prevent noncommunicable diseases and monitor healthcare costs [29, 28].

To bridge the gap between findings from controlled experiments and their implications for national policymaking, advanced causal inference methods for generalizing trial findings have been developed. In particular, generalizability frameworks enable estimation of robust population average treatment effects (PATE) by reweighting trial data to match the covariate distribution of a target population [32, 134, 41]. Weighting methods, such as logistic regression or entropy balancing, estimate the inverse-odds of trial participation (Lu et al.; see Ben-Michael et al. for discussion). Alternatives for estimating the PATE include g-methods, which model individual treatment effects directly using for example

linear regression or tree-based methods. Doubly robust estimators (also called augmented weighted estimators) [32], combine participation and outcome models.

Trial covariate adjustment is warranted due to the frequent divergence of participant characteristics in volunteer-based trials from those of the general population [135, 136]. Experimental samples are typically biased toward university students or individuals with elevated baseline health consciousness, education, and digital literacy [92]. This discrepancy stems from the sequential process of trial participation, which produces a sample whose distribution of effect-modifying characteristics substantially differs from that of the broader target population [137, 138]. When such characteristics govern individual responses to interventions, unadjusted extrapolation of trial results may substantially overestimate policy benefits [139, 92].

To accurately transport these experimental findings to the target population, researchers must estimate the PATE under the strict assumption that they have correctly identified a minimum separating set of covariates [140, 54]. This set includes the exact covariates needed to render the trial sampling mechanism conditionally independent of treatment effect heterogeneity. However, a pervasive methodological challenge arises because these essential covariates are often unavailable in both the experimental and observational datasets [54, 141, 142]. In practice, it is exceedingly difficult to ensure that all relevant treatment effect moderators have been fully captured, making missing covariates a central problem in generalizability frameworks. Omitting a critical effect modifier that differs in distribution between the source and target populations from this separating set violates the fundamental assumption of conditional exchangeability. This omission induces structural confounding that renders the resulting PATE estimates invalid for public policy modelling [143].

Within the doubly robust estimation framework, this structural bias arises from two main sources of error due to the omitted variable. First, it leads to residual covariate imbalance during selection, as the restricted sampling weights cannot fully adjust for non-random trial enrollment driven by the unobserved factor. Second, it produces residual treatment

effect heterogeneity, since the outcome model overlooks the moderating influence of the missing variable on individual responses to the intervention [144, 145].

To address this issue, the generalizability frameworks literature commonly employs parameter-based sensitivity analyses, which incorporate expert judgment into statistical formulations to quantify the expected magnitude and direction of bias from the missing covariate [54, 143]. By imputing or synthetically generating this covariate, these frameworks enable researchers to transparently justify plausible ranges for sensitivity parameters, rather than ignoring the omission. Although theoretically sound, they require specifying theoretical parameters and assuming particular data distributions to bound unmeasured confounding [143].

The application of these PATE estimators to nutrition-related data presents a persistent challenge. This difficulty arises because key mechanisms driving dietary choicesâ€"such as health consciousness and cognitive skillsâ€"are frequently unmeasured in target population surveys, thereby complicating robust generalizations. In the context of nutritional labelling, domain expertise posits working memory as a powerful treatment effect modifier [17, 112, 36]. An individual's ability to successfully utilize informational labels is strictly bounded by their cognitive capacity. Exceeding this limit induces more than mere comprehension failure; it actively triggers a health halo effect and moral licensing [146, 147]. Under cognitive overload [27], the sheer presence of nutritional data might grant consumers psychological permission to indulge, leading to suboptimal dietary choices and increased energy-dense intake [113]. Unfortunately, empirical measures of working memory or similar cognitive capacities are absent from nearly all national nutritional surveys and health records.

Utilizing data from a recent FOP label granularity trial [148] and the United Kingdom National Diet and Nutrition Survey as target population data, this study pursues two primary aims. First, we evaluate the methodological bias of omitting confounders by synthetically generating the unobserved working memory modifier when estimating PATE. To do this, we benchmark sequential hierarchical nested models relying on standard proxy

variables against a comprehensive separating set that includes the imputed cognitive variable. Specifically, causal effects are transported using a doubly robust estimator powered by Bayesian additive regression trees (BART) [149, 150]. The source trial is an online choice experiment designed to examine how label granularity influences consumer choices of breakfast cereals. The trial randomized participants to two granular labelling treatments with different cognitive demandsâ€"a coarse-label and a detailed-labelâ€"and measured working memory using n-back scores [148].

To accomplish these aim, this study advances the theoretical sensitivity paradigm for omitted confounders by moving beyond abstract statistical bounds [143]. Because working memory was fully measured in our source trial alongside standard demographic characteristics, we utilize generative machine learningâ€"specifically XGBoostâ€"to formally learn and synthetically reconstruct the unobserved working memory architecture for the NDNS population. This empirically grounded data generation allows us to rigorously test and explicitly isolate the structural bias in the PATE caused by the missing confounder. Furthermore, to defend this generative approach against imputation uncertainty, we systematically validate our findings by evaluating the PATE under deterministic extreme scenario boundsâ€"artificially perturbing the synthetically generated cognitive scores by $\pm 10\%$. This ensures that our final policy conclusions remain structurally stable regardless of the precise accuracy of the imputation.

Second, we incorporate these fixed PATE estimates into a dynamic ten-year microsimulation to rigorously assess the long-term public health burden and cost-effectiveness of each granular labelling counterfactual [151, 152, 153]. To enhance ecological validity, the microsimulation accounts for real-world behavioral constraints by modeling varying degrees of compensatory eating [154]. Moreover, it extends the evaluation beyond individual consumer behavior by examining the synergistic effects of combining label counterfactuals with mandatory industry reformulation [13]. Specifically, we investigate whether a label's cognitive complexity complements these structural benefits or if behavioral rebound actively counteracts reformulation gains.

Addressing this covariate omission is vital for accurate public policy forecasting. In contemporary health economic modeling, the internal validity of microsimulations hinges on the causal robustness of their input parameters [155, 152]. We argue that relying solely on standard demographic proxies to transport treatment effects produces biased PATE estimates, which carry over into subsequent health economic models. Incorporating these biased, proxy-based estimates into dynamic microsimulations distorts projected outcomes by masking potential adverse behavioral adaptations. This propagating error might misestimate upstream population BMI shifts that cascade into severe downstream clinical endpointsâ€"such as cardiovascular diseaseâ€"skewing projections of long-term noncommunicable disease burdens and associated healthcare costs. Ultimately, we demonstrate that parameter-based sensitivity analyses provides a crucial safeguard for policy evaluations, as the variance penalty from synthetic generation far outweighs the risks of ignoring structural cognitive heterogeneity.

The remainder of this paper is structured as follows. Section two establishes the causal framework, defining the potential outcomes notation, the structural assumptions required for transportability, and the theoretical challenge of the unobserved confounder. Section three details the empirical methodology, describing the trial and national survey data sources, the generative imputation approach, the doubly robust estimation strategy, and the design of the dynamic microsimulation. Section four presents our empirical findings, tracking the magnitude of the structural bias across nested models to forecast long term economic and health trajectories. Finally, section five discusses the policy and methodological implications of our work, addresses its limitations, and provides concluding remarks.

5.2 Methodology

5.2.1 Data sources and experimental design

The trial data come from a three-arm, between-subjects online experiment conducted to examine how the granularity of FOP labelling influences consumers' selection of breakfast cereals depending on their working memory capacity [148]. The trial's primary outcome was the average calorie count of cereals chosen by the treatment groups. Digital informed consent was obtained from all randomized participants, per University of Manchester Research Ethics Committee approval (Ref: 2024-16718-35972).

Prolific Academic platform recruited participants using quotas for sex, age, and ethnicity to mirror the UK adult census [39]. Of 570 initially screened individuals, 501 were randomized, and 498 remained for the final analytical sample (see Appendix). Subjects were assigned to one of three conditions: absent, coarse, or detailed. Granularity was manipulated via two calorie labels: a coarse, four-chunk format and a detailed, eight-chunk format, chosen to bracket the five-chunk Multiple Traffic Light standard familiar to UK shoppers [148].

FOP labels are designed to condense complex nutrient declarations into intuitive summaries. While prior work has examined colour and shape, far less attention has been paid to granularityâ€"the degree of detail, or number of informational chunks, contained in a label [8, 7]. Conceptually, granularity captures how finely a scheme partitions the healthfulness continuum. Cognitive-load theory posits that working memory can handle only a limited number of chunks at any moment; exceeding this capacity impairs comprehension and recall [27]. Classical estimates place this limit at roughly seven items [112], although contemporary resource models emphasize flexible allocation rather than a fixed span [113]. Consequently, a label that is optimal for a high-capacity consumer may overwhelm a shopper with lower capacity.

The present study addresses this by testing whether FOP label effectiveness can be optimized for consumers' working-memory capacity [148]. We anticipate a cognitive stratification pattern in which coarse labels promote healthier choices among individuals with lower working memory due to reduced cognitive demands, while detailed labels should benefit those with higher capacity by offering more comprehensive diagnostic information. Reframing label design as a problem of information-chunk optimization under resource constraints permits a direct examination of the label and cognitive-capacity interaction.

Participants' working memory was indexed with a three-stage n-back test [117, 118]. Letters appeared singly for 1 s each, with stimuli randomized across 1-back, 2-back, and 3-back blocks. Performance was scored as $d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$, where Z is the inverse cumulative normal. To ensure data quality in this unsupervised setting, participants needed to make fewer than 25 errors on the 1-back or 2-back task to advance, leading to final cognitive analytical samples of $n = 464$ and $n = 452$, respectively.

The target population was represented by data from the UK National Diet and Nutrition Survey (NDNS), a repeated cross-sectional study pooling data from 2008—2023 to derive population-level estimates [156]. We restricted the sample to adults aligned with the trial's eligibility criteria, yielding a final weighted target sample of 10,696 adults. To ensure the generalized estimates were representative of the UK population, all analyses of the target population data strictly incorporated the official NDNS survey weights.

5.2.2 Data harmonization and covariate selection

We harmonized a predefined array of covariates shared across the trial and NDNS datasets, including demographic variables (sex, age, and ethnicity) and socioeconomic variables (education level), which are recognized predictors of food selection and nutritional awareness [87, 89]. Household compositionâ€"particularly the presence of childrenâ€"was incorporated owing to its links with distinct purchasing behaviours and elevated nutritional priorities [85, 86].

Health metrics, including BMI and self-reported weight loss attempts, were also included as indicators of health consciousness, a factor apt to moderate responses to nutritional labeling [88]. Central to the analysis was cognitive performance (d'), which serves as a critical mediator in consumers' processing and utilization of intricate FOP label information [17, 148, 34, 36].

We distinguish between two distinct forms of missingness. Sporadic item non-response in the NDNS was imputed via Multiple Imputation by Chained Equations (MICE) [157]. The structural omission of d' scores in the NDNSâ€"absent from national surveys despite its theoretical importance as an effect modifierâ€"was addressed via the generative framework detailed in Section 3.4.

5.2.3 Identification and estimation framework

To bridge the divide between the trial sample and the target population, it is essential to differentiate between two interrelated concepts in the causal inference literature: generalizability and transportability [32]. Generalizability applies when the trial sample constitutes a probability-based random subset of the target population, enabling extrapolation of findings to that source population. In contrast, transportability is pertinent when the trial sample and target population are distinct and non-overlappingâ€"a scenario prevalent in public health research, where results from a volunteer-based experiment must be extrapolated to a separate national survey representing the broader population [32, 31].

To formalize this transportability analysis, this study adopts the Neyman—Rubin potential outcomes framework [40]. We conceptualize an infinite super-population from which all individuals are sampled [32]. Let $S \in \{0, 1\}$ be a random variable indicating study membership, where $S = 1$ denotes a participant sampled for the experimental RCT and $S = 0$ denotes an individual sampled for the target population, represented by the NDNS survey.

For each individual i in the super-population, let T_i be the treatment variable representing

the FOP labelling condition. Within the experimental trial ($S_i = 1$), T_i is randomly assigned to one of three levels: absent label ($T = a$), coarse label ($T = c$), or detailed label ($T = d$). Following the potential outcomes framework, we define $Y_i(t)$ as the potential outcome when unit i receives treatment $T_i = t$, where $t \in \{a, c, d\}$. The observed outcome, Y_i, represents the quantity of calories in the breakfast cereal selected by the individual.

Throughout the paper, we rely on the stable unit treatment value assumption (SUTVA), assuming no interference between units and that treatments are identically administered [40]. Assuming full compliance, we define the individual-level treatment effect, τ_i, as the explicit difference between the potential outcomes of unit i under contrasting label conditions: $\tau_i(t, t') = Y_i(t) - Y_i(t')$.

Because we can never simultaneously observe both potential outcomes for a specific unit, the individual-level treatment effect is inherently unidentifiable. We assume that both the trial sample $\{\tau_i, \mathbf{X}_i \mid S_i = 1\}_{i=1}^{N}$ and the target survey sample $\{\tau_i, \mathbf{X}_i, v_i \mid S_i = 0\}_{i=1}^{M}$ are drawn independent and identically distributed (i.i.d.) from the super-population, where $\mathbf{X}_i$ represents a vector of pre-treatment covariates and v_i represents the survey weight for the i-th individual in the NDNS [43, 143].

The sample average treatment effect (SATE) is defined as the average treatment effect restricted strictly to the experimental sample. Because treatment is randomly assigned with equal probability within the trial [143], a simple difference-in-means estimator can be used to recover the SATE:

$$\hat{\tau}\mathcal{S}(t, t') = \frac{1}{n_t} \sum i \in \mathcal{S} : T_i = t Y_i - \frac{1}{n_{t'}} \sum_{i \in \mathcal{S}: T_i = t'} Y_i, \tag{5.1}$$

where $\mathcal{S}$ represents the set of indices corresponding to units in the experimental sample, and n_t and $n_{t'}$ represent the number of units randomly assigned to treatments t and t'.

However, the true population average treatment effect (PATE) is our causal quantity of interest. The true PATE represents the theoretical mean (expectation) of the treatment

effects if the entire target population were exposed to the intervention, formally defined as:

$$\tau \equiv \mathbb{E}[\tau_i \mid S_i = 0] \tag{5.2}$$

If the experimental sample were a perfect random draw from the super-population, the empirical SATE ($\hat{\tau}_{\mathcal{S}}$) would be an unbiased estimator for the true PATE. Yet, online volunteer trials inherently suffer from covariate shiftâ€"systematic differences in the distributions of demographic and cognitive characteristics between the trial participants and the general target population. Consequently, SATE $\neq$ PATE, and experimental results cannot be naively extrapolated [54, 143].

To recover the true PATE from a biased experimental sample, we must rely on three core structural assumptions [32, 158]. First, we assume internal ignorability. Within the trial, treatment assignment is randomized and therefore independent of potential outcomes, denoted as $Y_i(t) \perp T_i \mid S_i = 1$. Second, we require positivity or overlap. For any covariate profile $\mathbf{x}$ present in the target population, the probability of trial participation must be strictly positive, expressed as $\Pr(S_i = 1 \mid \mathbf{X}_i = \mathbf{x}) > 0$.

Third, we assume conditional exchangeability, which requires identifying a separating set of covariates [140]. This set comprises pre-treatment variables that fully account for non-random selection into the trial and block all confounding between study participation and treatment effect heterogeneity. We denote this set by $\mathbf{X}_i$, under the assumption that it renders treatment effects τ_i independent of study membership S_i conditional on $\mathbf{X}_i$: $\tau_i \perp S_i \mid \mathbf{X}_i$.

When these assumptions hold, researchers traditionally estimate the PATE using a weighted estimator that reweights trial observations to match the target population's covariate distribution [41]. In practice, this involves up-weighting trial participants whose cognitive and demographic characteristics are underrepresented relative to the national

survey, while down-weighting those who are overrepresented. The sampling weight for each unit i in the experimental sample is formally defined as the inverse odds of sampling:

$$w_i = \frac{\Pr(S_i = 1)}{\Pr(S_i = 0)} \cdot \frac{1 - \Pr(S_i = 1 \mid \mathbf{X}_i)}{\Pr(S_i = 1 \mid \mathbf{X}_i)} \tag{5.3}$$

where the first term represents the marginal odds of selection into the target population, and the second term is the inverse of the conditional odds of trial participation given the separating set $\mathbf{X}_i$. These weights are derived from a dedicated sampling model.

5.2.3.1 Doubly robust estimation strategy: TMLE-BART

To maximize estimation robustness, this study employs a doubly robust (DR) estimator, also known as an augmented weighted estimator [32]. This framework is semi-parametrically efficient and allows practitioners to simultaneously leverage both a sampling model and an outcome model for the individual-level treatment effect. We employ the Targeted Maximum Likelihood Estimation (TMLE) framework [159] powered by the Bayesian Additive Regression Tree (BART) algorithm [48, 50] to estimate the nuisance components. The final estimator for the PATE between conditions t and t' is given by:

$$\hat{\tau} W^{DR}(t, t') = \hat{\tau}\mathcal{W}(t, t') - \frac{1}{N} \sum_{i:S_i=1} w_i \hat{\tau}(t, t', \mathbf{x}i) + \frac{1}{\sum j : S_j = 0 v_j} \sum_{j:S_j=0} v_j \hat{\tau}(t, t', \mathbf{x}_j) \tag{5.4}$$

where $\hat{\tau}_{\mathcal{W}}$ denotes the conventional weighted estimator; the intermediate sum aggregates over the N units in the experimental sample ($S_i = 1$); and the final sum aggregates over the M units in the target survey sample ($S_j = 0$), incorporating the NDNS survey weights v_j.

This formulation establishes a robust safeguard by integrating sampling weights with outcome model predictions. It derives a bias-correction term by subtracting the predicted

weighted outcomes from the observed weighted outcomes in the trial, thereby isolating the outcome model's specification error. This error is then incorporated into the outcome model's predictions for the target population, weighted by the survey design. This structure confers double robustness, ensuring the PATE estimator remains consistent and asymptotically unbiased if either the outcome model or the sampling model is correctly specified [43, 50]. BART organically captures non-linear interactions without manual specification and incorporates NDNS survey weights during the prediction phase for the target survey sample.

A mathematical consequence of integrating synthetically generated covariates into the estimation pipeline is the variance penalty. Unlike observed demographic variables, the imputed cognitive capacity carries inherent estimation uncertainty that inflates the standard errors of the outcome model. This variance penalty represents the increase in estimator variance that must be tolerated to mitigate structural bias; we acknowledge this trade-off as a necessary price for the causal identification of the target effect. Models were implemented in R using the `dbarts` package with 200 trees and 1000 post-burn-in MCMC iterations.

5.2.4 Addressing unobserved confounding: generative benchmarking

Researchers typically estimate the PATE under the strict assumption that they have correctly identified the minimum separating set. When this exact set of covariates is fully observed, DR estimators consistently recover the target effect. However, a pervasive problem arises in applied transportability analysis due to severe data constraints. While experimental trials often collect rich behavioral and cognitive data, national observational surveys are typically limited to standard demographic variables. Consequently, a key unobserved variable, U_i (working memory), is frequently measured in the source data but completely absent from the target population data [54].

To begin, we formally define the minimum separating set as $\mathcal{X}_i = \mathbf{X}_{obs,i}, U_i$, where $\mathbf{X}_{obs,i}$ is observed, and U_i is not. For the DR estimator to be unbiased, both components would have to be included in the sampling weights and the outcome model; however, due to data constraints, we omit U_i. When U_i is omitted, the exchangeability assumption is severely violated Figure Fig. 5.1. If researchers naively estimate the transported effect using only these restricted covariates, both models become misspecified, yielding an unadjusted and biased estimator:

$$\hat{\tau}W, mis^{DR}(t,t') = \hat{\tau}\mathcal{W}, mis(t,t') - \frac{1}{N}\sum_{i:S_i=1} w_i\hat{\tau}(t,t',\mathbf{x}obs,i) + \frac{1}{\sum j : S_j = 0 v_j}\sum_{j:S_j=0} v_j\hat{\tau}(t,t',\mathbf{x}_{obs,j}) \tag{5.5}$$

This misspecified estimator yields a biased estimation of the PATE. Colnet et al. [54] mathematically formalize the expected structural bias induced by this omission, denoted as $\mathcal{B}$, as the explicit difference between the misspecified estimate and the true, structurally complete DR target effect $\hat{\tau}_W^{DR}(t,t')$:

$$\mathcal{B} = \hat{\tau}W, mis^{DR}(t,t') - \hat{\tau}W^{DR}(t,t') \tag{5.6}$$

The structural bias $\mathcal{B}$ arises from two primary sources of error. First, it reflects residual covariate imbalance in the selection process, as the constrained sampling weights inadequately adjust for non-random trial enrollment influenced by the unobserved factor. Second, it captures residual treatment effect heterogeneity, since the restricted outcome model overlooks the moderating role of the missing variable on individual intervention responses. Following contemporary developments, we assess the robustness of transported estimates by synthetically generating the missing values [143, 142](see Appendix).

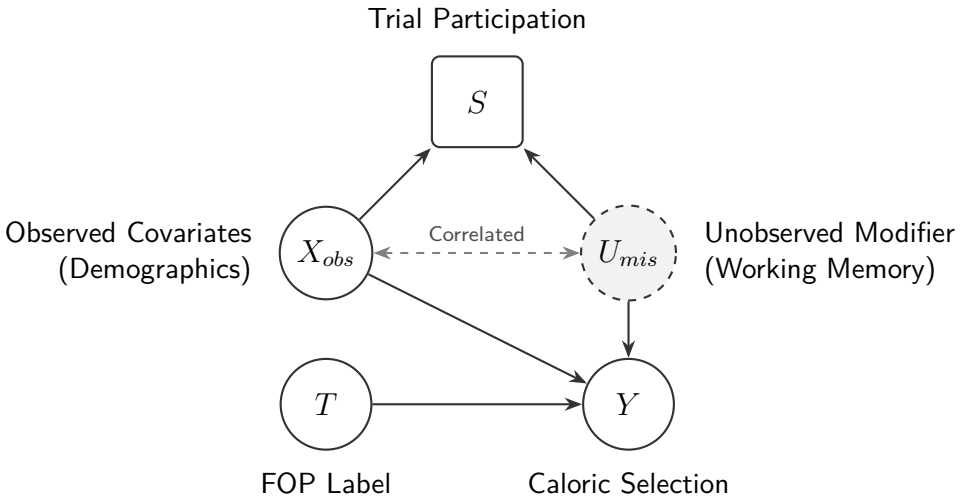

Fig. 5.1. Directed acyclic graph illustrating the structural problem of unmeasured confounding in transportability analysis. Schematic representation adapting the causal framework of [54]. The sampling indicator (S) is jointly determined by universally observed covariates (X_{obs}) and the unobserved cognitive modifier (U_{mis}). Within the experimental trial ($S = 1$), the randomly assigned treatment (T) and baseline covariates jointly influence the behavioral outcome (Y). Because U_{mis} dictates both the probability of trial participation and individual treatment effect heterogeneity, omitting it from the adjustment separating set strictly violates the conditional mean exchangeability assumption. This omission induces a structural bias when extrapolating the trial findings to the target population.

5.2.4.1 Generative imputation via XGBoost

We deploy generative machine learningâ€"specifically XGBoost (Extreme Gradient Boosting)â€"to formally learn the non-linear relationship $f(\mathbf{X}_{obs}) \rightarrow U$ within the trial sample and synthetically reconstruct the missing working memory architecture for the target survey [160, 161]. XGBoost was selected for its ability to capture high-order interactions between demographics and cognitive performance without manual parameterization [162].

5.2.4.2 Nested modeling trajectory and deterministic scenario bounds

To empirically quantify and visualize the structural bias $\mathcal{B}$ within our specific context, we estimate the PATE across sequential hierarchical nested models. By explicitly calculating

the difference between proxy-reliant baseline models and our fully benchmarked model that conditions on the synthesized variable, we isolate the exact magnitude of the bias.

To defend this generative approach against imputation uncertainty and satisfy the variance penalty associated with synthetic generation, we perturb our benchmarked estimates by deterministic boundaries shifted by $+10\%$ and -10%. This ensures the structural stability of our final policy effect is rigorously evaluated against unmeasured confounding without relying on abstract mathematical parameters [143, 44]. We argue this variance penalty is mathematically superior to accepting the severe structural bias $\mathcal{B}$ of a model that omits cognitive heterogeneity.

5.2.5 Long-term forecasting: dynamic microsimulation

We incorporate the fixed PATE estimates into a dynamic, stochastic, individual-level microsimulation [151, 153]. This engine operates as a discrete time state transition model, modeling the natural history of cardiometabolic health through annual update cycles [20]. The baseline virtual cohort was constructed using individual level records from the NDNS spanning 2008 to 2023. To preserve national representativeness across this 16-year structure, we implemented temporal pooling, assigning weights based on the duration of each survey wave relative to the total study period. All analyses of this cohort strictly utilized official survey weights to ensure population fidelity. The final virtual cohort was then scaled using a national multiplier to represent the 53 million adults in the United Kingdom [163] (see Appendix).

5.2.5.1 Dynamic energy balance model

To ensure biological plausibility, longitudinal BMI transitions were modeled using validated dynamic energy balance equations [164]. We adopted a dynamic physiological framework where a persistent reduction of 55 kcal per day results in a body weight reduction of 0.4536 kg over approximately 1 year, with 50% of the change occurring within the first 12

months. The simulated transition follows an exponential decay toward a new metabolic equilibrium, strictly defined by an adaptation rate of 1 kg of weight change per 24 kcal per day of persistent dietary shift [20].

To further reflect behavioral reality in free living environments, we implemented a dietary compensation parameter, denoted as γ. This parameter represents the psychological manifestation of moral licensing and the health halo effect [146], where perceived healthiness triggers compensatory consumption that dilutes the raw policy caloric shock, Δ_{kcal}, such that the net biological impact is mathematically defined as:

$$\Delta_{net} = \Delta_{kcal} \times (1 - \gamma) \tag{5.7}$$

Our primary models assumed a central compensation parameter of 0.5, dictating that exactly half of the initial caloric deficit is subsequently offset by compensatory eating [151]. We conducted rigorous sensitivity analyses around this behavioral parameter using bounds of 0.3 and 0.7.

5.2.5.2 Labelling counterfactual scenarios

The simulation evaluated four primary counterfactual scenarios to disentangle the impact of consumer behavior from industry level structural changes [13, 19]. The first two scenarios evaluated the consumer only responses to the coarse label and the detailed label, utilizing the DR point estimates derived from our causal transportability models. The subsequent two scenarios extended these behavioral frameworks by incorporating mandatory industry reformulation.

To assess the additional impact of potential product reformulation as a supply-side response to granular FOP labeling, we estimated a response corresponding to the total net reduction in the caloric density of breakfast cereals over the simulated period. Drawing from the framework of Liu et al. [151] and evidence from successful UK nutrient reduction programsâ€"specifically the Public Health England sugar reduction strategy [165] and

the salt reduction initiative [166]â€"we evaluated a modest 5% net calorie reduction in cereals as a result of the policy implementation. We assumed this reformulation effect would occur in a staged manner from the first year through year 5 of the intervention (i.e., 1% per year for 5 years), with no additional reformulation thereafter.

The model incorporates a survival engine to prevent immortal time bias, ensuring that simulated individuals who die cease contributing to population level health and economic aggregates. Baseline mortality risks were derived from the 2024 Office for National Statistics death registrations for England and Wales [167], stratified by single year age and sex. These baseline hazards, $h_0(a, s)$, were dynamically adjusted in each annual cycle using BMI-specific risk factors [168]. Mortality risk was adjusted log linearly for individuals with a BMI above the nadir of 25 kg/m^2, applying a hazard ratio of 1.21 per 5 kg/m^2 increase:

$$h_i(t) = h_0(a, s) \times 1.21^{\frac{BMI_i(t)-25}{5}} \tag{5.8}$$

A survival step was performed annually, where individuals were marked as deceased if a randomly drawn uniform variable fell below their adjusted probability of death. To isolate the causal effect of the labeling policies from stochastic biological noise, we utilized common random numbers. The matrices of uniform random death draws and autoregressive biological weight fluctuations [169], parameterized with an autocorrelation coefficient of 0.85 and a shock variance of 0.05, were pre generated and identically applied across all counterfactual arms.

5.2.5.3 Health and economic outcomes

Health and economic outcomes were evaluated over a 10-year horizon from a societal perspective, specifically tracking cumulative changes in daily calories, absolute BMI shifts, obesity prevalence, quality-adjusted life years (QALYs), and healthcare expenditures [151, 153]. Direct healthcare burdens to the National Health Service were estimated at £16 per excess BMI unit per person year [170], while health utilities were modeled as a loss of

0.003 QALYs per BMI unit increase [171].

All future costs and health gains were discounted at an annual rate of 1.5%, consistent with HM Treasury Green Book guidelines for long term health valuations [172]. Probabilistic sensitivity analysis was executed utilizing 1,000 Monte Carlo iterations [152], simultaneously sampling uncertainty in the causal intervention effect alongside random probability draws for health economic parameters. Final results were reported as medians and 95% uncertainty intervals derived from the 2.5th and 97.5th percentiles.

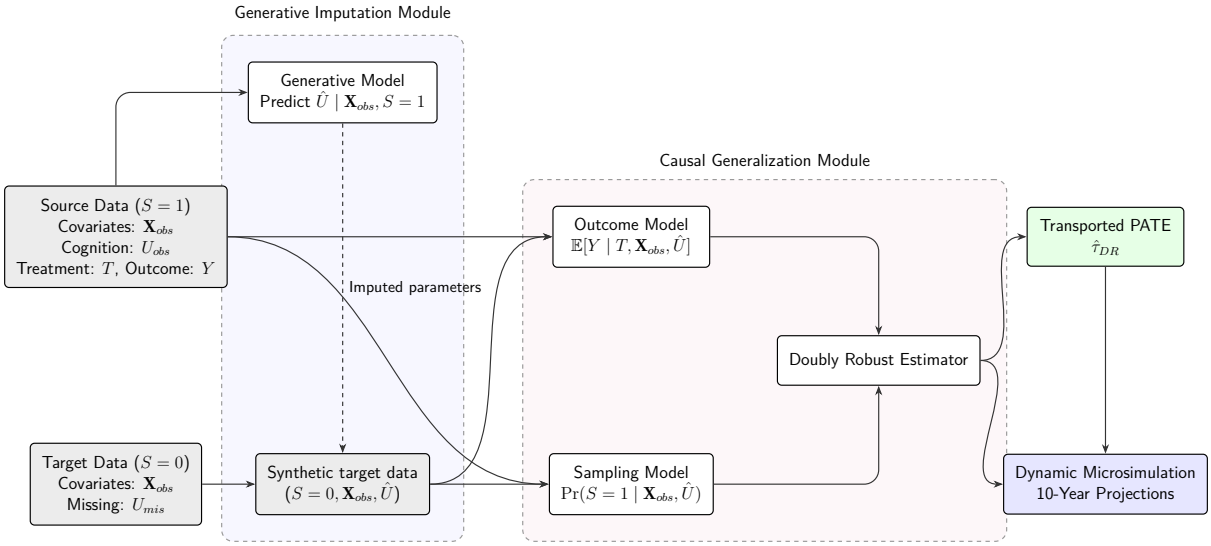

Fig. 5.2. Generative imputation process via XGBoost. $S = 1$ denotes the source experimental sample, and $S = 0$ denotes the target national population. $\mathbf{X}_{obs}$ represents the vector of universally observed pre-treatment covariates. U_{obs} represents the working memory scores fully observed within the trial, while U_{mis} denotes the unmeasured cognitive capacity in the target survey. $\hat{U}$ signifies the synthetically imputed cognitive scores. T indicates the assigned label treatment, Y denotes the caloric choice outcome, and $\hat{\tau}_{DR}$ represents the doubly robust estimator for the population average treatment effect.

Table 5.1. Baseline characteristics of the randomized control trial and survey-weighted National Diet and Nutritional Survey (2008-2023) populations baseline.

Characteristic	Trial (N=498)	NDNS (N=10696)
Sex		
Male	240 (48.2%)	5190 (48.5%)
Female	258 (51.8%)	5506 (51.5%)
Age Group		
18-24	56 (11.2%)	1146 (10.7%)
25-34	84 (16.9%)	1754 (16.4%)
35-44	84 (16.9%)	1919 (18.0%)
45-54	81 (16.3%)	1853 (17.3%)
55-64	141 (28.3%)	1614 (15.1%)
65 and over	52 (10.4%)	2410 (22.5%)
Ethnicity		
Asian or Asian British	37 (7.4%)	47.2 (0.0%)
Black, Black British, Caribbean or African	20 (4.0%)	0 (0.0%)
Mixed or multiple ethnic groups	12 (2.4%)	1779 (16.6%)
Other ethnic group	7 (1.4%)	0 (0.0%)
White	422 (84.7%)	8870 (83.0%)
Education		
Higher Education	101 (20.3%)	2270 (21.2%)
College or university	330 (66.3%)	4944 (46.2%)
Secondary school	64 (12.9%)	1680 (15.7%)
Primary school	1 (0.2%)	1802 (16.8%)
Prefer not to say	1 (0.2%)	0 (0.0%)
Children in HH		
No	340 (68.3%)	7281 (68.1%)
Yes	158 (31.7%)	3415 (31.9%)
Trying to Lose Weight		
No	229 (46.0%)	8938 (83.6%)
Yes	269 (54.0%)	1758 (16.4%)
BMI (kg/m^2)	28.83 (26.26)	27.47 (5.53)
Working Memory (2-back d')	3.21 (0.95)	2.74 (1.02)
Working Memory (3-back d')	2.25 (0.86)	1.78 (0.97)

Note: Baseline sociodemographic and cognitive characteristics of the trial and target (NDNS) Populations. Categorical variables are presented as n (%), and continuous variables as mean (SD). NDNS estimates incorporate survey sampling weights to reflect the broader UK adult population. Missing baseline covariates in the NDNS dataset were imputed using Multiple Imputation by Chained Equations (MICE). Working memory capacities (2-back and 3-back d') for the NDNS cohort denote synthetically generated parameters.

5.3 Results

5.3.1 Baseline characteristics and covariate balance

Table Table 5.1 details the baseline sociodemographic, behavioral, and cognitive characteristics of the unadjusted randomized controlled trial sample (N=498) compared to the survey-weighted UK target population derived from the NDNS (N=10,696). While randomization ensures internal validity across the trial arms, a direct comparison against the NDNS target population reveals notable baseline discrepancies, underscoring the presence of a healthy volunteer selection bias. The proportion of participants actively trying to lose weight is more than three times higher in the trial (54.0%) compared to the general population (16.4%). Similarly, the mean BMI in the trial sample (28.83 kg/m^2) is higher than the population average (27.47 kg/m^2), underscoring that while RCTs establish efficacy within a specific context, they may not accurately quantify the likely impact of the policy across a diverse nationwide population [43].

Certain structural and demographic variables exhibit highly comparable distributions across both datasets. For instance, the gender distribution is exceptionally well-balanced, with the trial comprising 48.2% males and 51.8% females, closely mirroring the NDNS population (48.5% and 51.5%, respectively). Similarly, the proportion of individuals residing in households with children is nearly identical (31.7% in the trial versus 31.9% in the target population). However, the age distribution demonstrates structural divergence. The late-middle-aged bracket (55—64 years) is substantially overrepresented in the trial (28.3% versus 15.1% in the target population), whereas the elderly population (aged 65 and over) is heavily underrepresented, making up only 10.4% of the trial sample compared to 22.5% nationally.

Pronounced disparities are also evident across socioeconomic and ethnic indicators. The trial cohort is disproportionately highly educated, with 86.6% of participants holding a higher education or college/university degree, compared to 67.4% in the general population.

Conversely, individuals whose highest attainment is primary school are virtually absent from the trial (0.2%) but constitute a meaningful fraction of the national target (16.8%). In terms of ethnic composition, while the White majority is comparable (84.7% in the trial versus 83.0% in the NDNS), the representation of minority ethnic groups differs, with the trial capturing a higher proportion of Asian or Asian British participants (7.4%) but fewer individuals from mixed or multiple ethnic backgrounds (2.4% versus 16.6%).

Finally, substantial differences were observed in cognitive performance, confirming the selection bias in the domain of working memory. Trial participants displayed significantly higher scores compared to the semi-synthetic estimates generated for the target population on both the standard 2-back task (3.21 vs 2.74) and the highly demanding 3-back task (2.25 vs 1.78). Collectively, these descriptive statistics confirm that a naive generalization of the trial data would inadvertently project policy impacts onto a population characterized by a fundamentally different demographic, behavioral, and cognitive profile than the actual population of interest.

5.3.2 Covariate balance

To resolve these representativeness failures, the inverse probability weights were calculated to map the trial respondents onto the target demographic profile. Figure Fig. 5.3 A) visually confirms the efficacy of this transportability weighting procedure. The plot displays the Absolute Standardized Mean Differences (ASMD) for key covariates before and after applying the weights. As indicated by the hollow black circles, the unadjusted naive trial sample exhibited severe baseline imbalances, heavily exceeding the standard diagnostic threshold of 0.1. Disparities were particularly extreme for weight loss intentions (ASMD > 0.8), college education, and specific age brackets. However, the application of the transportability weights (solid grey circles) systematically collapsed these discrepancies. Following adjustment, the ASMD for all modeled covariates fell well below the 0.1 threshold of negligible imbalance, mathematically confirming the successful reconstruction of the

NDNS target population's covariate profile within the trial data.

Figure Fig. 5.3B) illustrates the downstream consequences of achieving this covariate balance on the estimated intervention outcomes. The forest plot contrasts the unadjusted naive trial effects against the transported Population Average Treatment Effects (PATE) across the nested model specifications (Models A through F). The data reveal that adjusting for covariates fundamentally alters the predicted policy efficacy. For example, while the unadjusted naive estimates suggest a specific magnitude of caloric count reduction, the transported point estimates systematically shiftâ€"and in some specifications, reverse directionâ€"once the target population's true distribution of age, education, and cognitive capacity is accounted for. Furthermore, the varying point sizes (scaled to Confidence Interval width) highlight the dynamic uncertainty introduced as the models become more highly specified. Ultimately, these parallel plots demonstrate that evaluating labelling policies through a naive, unweighted lens fails to capture the true, localized impact those interventions will have when deployed at a national scale.

5.3.3 Population Average Treatment Effects (PATE)

The primary analysis evaluated transported PATE across a series of nested models to quantify the impact of unmeasured cognitive confounding. Table Table 5.2 presents the comparative analysis between trial benchmarks, derived via unadjusted and interaction-based ordinary least squares (OLS) regressions, and PATE estimates transported via the DR TMLE-BART framework.

The results show a divergence between efficacy observed within the restricted trial setting and projected impact on the general NDNS population. In the naive trial analysis, the estimated main effects of the nutritional label on caloric count reduction were -9.1 kcal [95% CI: -23.4, 5.1] for the coarse group and -3.5 kcal [95% CI: -18.1, 11.2] for the detailed group when compared to the absent condition.

The regression interaction analysis show negative interaction coefficients between the

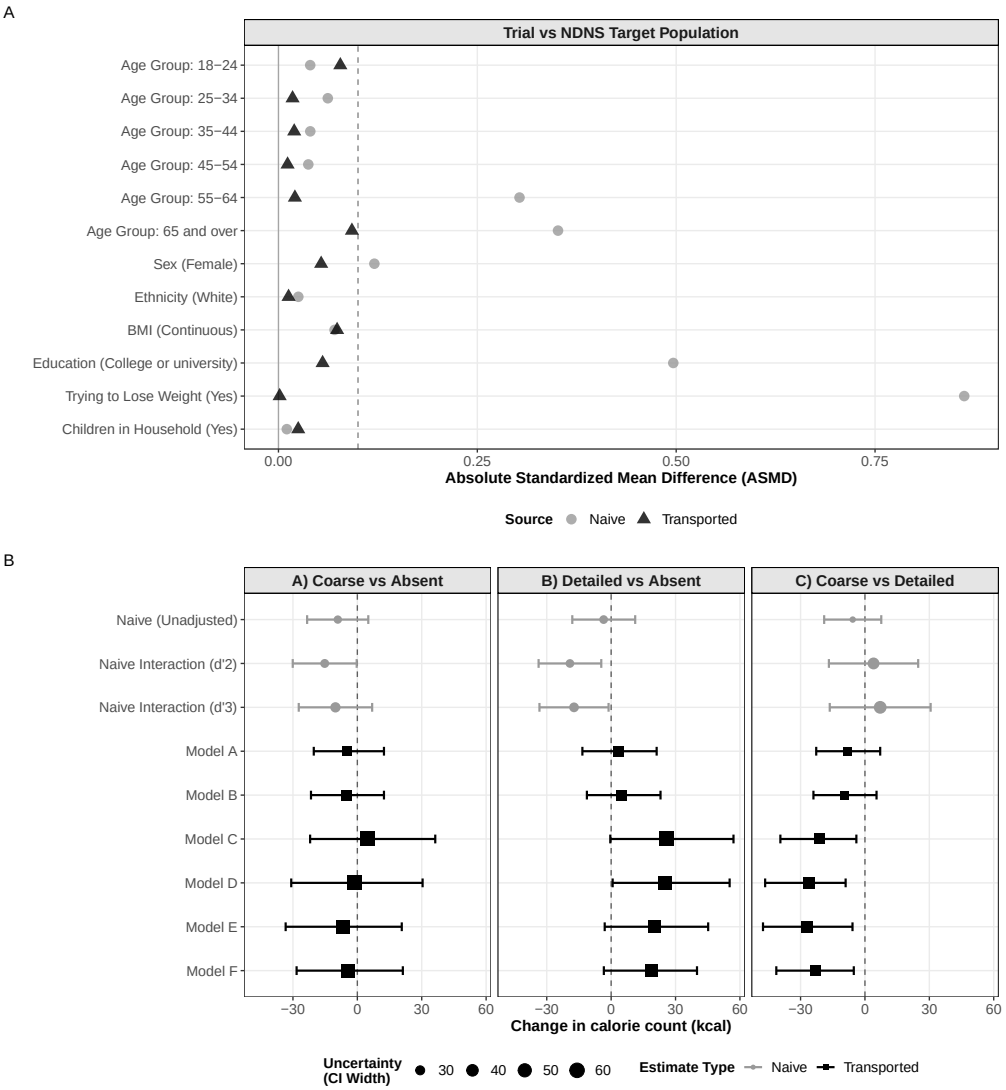

Fig. 5.3. Covariate balance and transported population average treatment effects (PATE) via logistic regression. (A) Covariate balance between the pooled trial sample and the NDNS target population before and after transportability weighting. The plot displays the Absolute Standardized Mean Differences (ASMD) for key demographic, lifestyle, and socioeconomic covariates. Hollow black circles (Naive) represent the baseline imbalance between the unadjusted trial sample and the survey-weighted NDNS population. Solid grey circles (Transported) represent the balance achieved after applying the inverse probability weights. The vertical dashed line at 0.1 represents the standard threshold for negligible imbalance; values to the left of this line indicate successful reconstruction of the target population's covariate profile within the trial sample. (B) Forest plot of the estimated treatment effects on caloric count. The plot displays the naive trial estimates versus the transported PATEs across three comparisons: Coarse vs. Absent, Detailed vs. Absent, and Coarse vs. Detailed. Point sizes are scaled to the width of the 95% Confidence Intervals to illustrate estimation uncertainty.

detailed label and continuous working memory scores across the 2-back task (-19.2 kcal [95% CI: -33.8, -4.6]) and the 3-back task (-17.3 kcal [95% CI: -33.4, -1.2]).

When these estimates are transported to the NDNS population using the DR estimator, the pattern of policy effectiveness shifts. Model A, which adjusts strictly for demographic disparities, yields non-significant estimates. The inclusion of household, lifestyle, and socioeconomic variables in Models B, C, and D results in a divergence in the predicted efficacy. By Model C, which adjusts for BMI and intent to lose weight, the coarse label results in lower calorie choices compared to the detailed label, with a relative reduction of -21.1 kcal [95% CI: -39.4, -4.0].

The integration of socioeconomic constraints in Model D and synthetically generated cognitive measures in Models E and F provides evidence of selection bias. Adjusting for socioeconomic status in Model D results in the detailed label increasing calorie count by 25.0 kcal [95% CI: 0.7, 55.2] relative to the absent group. When accounting for the average working memory of the general population in Model E, the point estimate remains at 20.0 kcal [-3.0, 45.2]. In Model E, the coarse label reduces intake by 26.9 kcal [95% CI: -47.5, -5.8] relative to the detailed label, and -22.9 kcal [95% CI: -41.3, -5.2] in Model F.

The difference in PATE estimates between Model A and the cognitively adjusted Model E shows an explicit 18.9 kcal shift for the coarse versus detailed comparison. This shift quantifies the structural bias induced by unmeasured cognitive confounding.

The variation from Model A through Model F shows the trade-off between bias reduction and variance inflation inherent in this causal transportability framework Fig. 5.3B). The naive trial estimates, represented by the unadjusted points, exhibit artificially high precision due to their narrow confidence intervals, yet they systematically fail to account for the population's distinct socioeconomic and cognitive profile. As the transportability models incorporate increasingly covariates from Model A through Model F, there is a substantial leftward and rightward shifting of the point estimates, structurally correcting for the selection bias. However, this rigorous mathematical adjustment incurs a necessary

cost in precision. Notably, while Models D, E, and F introduce the greatest statistical uncertainty, they represent the only specifications that explicitly adjust for the critical socioeconomic and cognitive effect modifiers identified during the trial phase. This indicates that these statistically wider estimates reflect an ecologically valid approximation of the true population-level effects that policymakers can expect upon national rollout.

5.3.4 Health and economic outcomes

The 10-year dynamic microsimulation reveals that the long-term public health trajectory of the UK is highly sensitive to the interaction between label granularity and the complexity of the transportability adjustment. Under the status quo counterfactual representing natural weight gain, the virtual population experiences a cumulative NHS burden starting at £11.5 million in Y1 and rising to £325.0 million by Y10.

In scenarios relying solely on consumer behavior, informational complexity is a primary driver of health deficits in the general population. The progression from the demographic-only Model A to the fully cognitively adjusted Model F illustrates a severe divergence in policy efficacy. Under Model D, the detailed label predicts highly deleterious outcomes, triggering a cognitive backfire that results in 392,900 additional obesity cases, a cumulative loss of 306,400 QALYs, and £1.15 billion in excess NHS costs by Y10. Variation across the transportability adjustments is stark: for the detailed label scenario, Model A projects only 50.3k additional obesity cases, whereas Model F projects 294.0k incident cases. This discrepancy of 243.7k obesity cases reflects the impact of omitting working memory, the true confounder, from the separating set. In Model E, which accounts for moderate-difficulty working memory ($d'2$), the detailed label generates 262,100 [95% UI: 181,100, 354,800] additional obesity cases in Y1, rising to 316,300 [95% UI: 222,900, 420,000] by Y10. Economically, this detailed signaling incurs £48.1 million [95% UI: £36.6, £58.4] in incremental NHS costs in Y1, escalating to £922.6 million [95% UI: £701.5, £1141.0] by Y10.

Table 5.2. Comparison of FOP labelling effects on caloric count: Randomized controlled trial benchmarks from OLS models versus transported Population Average Treatment Effects (PATE) via doubly robust TMLE-BART across nested models.

Model	N_{RCT}	N_{NDNS}	*Coarse vs Absent*	$\mathcal{B}$	*Detailed vs Absent*	$\mathcal{B}$	*Coarse vs Detailed*	$\mathcal{B}$
Naive (OLS Unadjusted)								
FOP labelling	490	–	-9.1 [-23.4, 5.1]	–	-3.5 [-18.1, 11.2]	–	-5.7 [-19.0, 7.6]	–
Naive (OLS Adjusted - Interaction effects)								
FOP labelling*($d'2$)	464	–	-15.2 [-30.1, -0.3]	–	-19.2 [-33.8, -4.6]	–	4.0 [-16.8, 24.8]	–
FOP labelling*($d'3$)	452	–	-10.2 [-27.3, 6.9]	–	-17.3 [-33.4, -1.2]	–	7.1 [-16.4, 30.6]	–
Transported PATE (Doubly Robust TMLE-BART)								
Model A (Demographics)	498	10,696	-4.6 [-20.3, 12.4]	0.0	3.4 [-13.4, 21.2]	0.0	-8.0 [-22.7, 7.1]	0.0
Model B (+ Household)	498	10,696	-4.9 [-21.6, 12.4]	0.3	4.6 [-11.3, 23.0]	-1.2	-9.5 [-24.0, 5.4]	1.5
Model C (+ Lifestyle)	498	10,696	4.8 [-22.0, 36.3]	-9.4	25.9 [-0.4, 57.0]	-22.5	-21.1 [-39.4, -4.0]	13.1
Model D (+ SES)	498	10,696	-1.1 [-30.8, 30.4]	-3.5	25.0 [0.7, 55.2]	-21.6	-26.0 [-46.5, -9.0]	18.0
Model E (+ $d'2$)	464	10,696	-6.9 [-33.4, 20.7]	2.3	20.0 [-3.0, 45.2]	-16.6	-26.9 [-47.5, -5.8]	18.9
Model F (+ $d'3$)	452	10,696	-4.3 [-28.3, 21.2]	-0.3	18.6 [-3.4, 40.0]	-15.2	-22.9 [-41.3, -5.2]	14.9

Note: Naive (OLS Unadjusted): Average treatment effects (ATE) of FOP labelling on calorie count from a randomized controlled trial (RCT) reported in Avalos [148]. *Naive (OLS Adjusted):* Interaction effects for FOP labelling comparisons evaluated at working memory capacities (d') from the RCT dataset. *Doubly Robust TMLE-BART:* Population average treatment effects (PATE) for the NDNS population. The Bias ($\mathcal{B}$) column represents the absolute subtraction of the specific comparison estimate from the Model A baseline ($Estimate_{ModelA} - Estimate_{Modeli}$), quantifying the shift induced by incremental covariate adjustment. Model specifications: *Model A:* Age, sex, ethnicity. *Model B:* Model A + children in household. *Model C:* Model B + BMI, intent to lose weight. *Model D:* Model C + education level. *Model E/F:* Model D + generative synthetic working memory (d'_2 / d'_3) using XGBoost.

Conversely, Model E represents the setting with the greatest health benefits and lowest costs under the coarse label scenario. This label prevents 86,200 [95% UI: 45,900, 141,800] obesity cases in Y1, rising to 105,900 [95% UI: 56,100, 172,300] by Y10. This yields a gain of 3,200 [95% UI: 2,100, 4,300] QALYs in Y1 and stabilizes at a gain of 79,300 [95% UI: 35,600, 166,400] QALYs by Y10. Under Model F, the coarse label prevented 64.5k obesity cases and gained 45.7k QALYs. Net savings to the NHS expand from £16.5 million [95% UI: £12.3, £20.3] at Y1 to £316.5 million [95% UI: £231.1, £398.7] by the tenth year.

The integration of mandatory industry reformulation reconfigures the policy hierarchy by providing a structural caloric deficit that competes with consumer-level licensing effects. The synergistic strategy combining coarse labels with industry reformulation (Scenario C) produces the most favorable overall outcomes, compounding the peak consumer benefits observed in the cognitively adjusted models. In Model F, Scenario C prevents 439.9k obesity cases, gains 186.4k QALYs, and results in £765.1 million in NHS savings. In Scenario C under Model E, obesity preventions scale from 119,600 [95% UI: 67,000, 183,100] in Y1 to 473,600 [95% UI: 356,500, 610,600] by Y10. This translates to an initial gain of 4,300 [95% UI: 2,800, 5,700] QALYs in Y1, accelerating to 219,100 QALYs [95% UI: 127,800, 338,000] by Y10. Net fiscal savings to the NHS expand from £22.5 million [95% UI: £16.7, £27.9] at Y1 to a total of £885.3 million [95% UI: £650.1, £1106.8] at Y10.

While industry reformulation partially mitigates the backfire of the detailed label (Scenario D), this combination remains decidedly less effective than coarse approaches. In Scenario D under Model F, the complex label's behavioral rebound partially offsets reformulation benefits, yet prevents 79.4k obesity cases by Y10; however, this results in a net loss of 82.5k QALYs and £291.2 million in additional NHS expenditures. In Model E, caloric surpluses result in 228,500 [95% UI: 151,300, 320,500] additional obesity cases in Y1, though accumulated structural savings eventually result in 59,100 [95% UI: 24,600, 107,600] cases prevented by Y10. Despite this late-term physiological shift, the cumulative health impact remains negative throughout the decade, with a loss of 8,000 [95% UI: 5,600,

10,800] QALYs in Y1 progressing to a net loss of 101,700 QALYs [95% UI: 36,700, 206,700] by Y10. Total incremental costs to the NHS increase from £42.0 million [95% UI: £31.7, £52.7] in Y1 to £353.8 million [95% UI: £260.3, £455.2] by Y10.

The robustness of these divergent trajectories is visually confirmed in the multi-panel trend plots for absolute BMI (Figure Fig. 5.5), incremental QALYs (Figure Fig. 5.4), and NHS costs (Figure Fig. 5.6). These figures stratify simulated outcomes across three assumed levels of dietary compensation (30%, 50%, and 70%) and the sequential progression of the transportability models (Model A through F). Across all compensation behaviors, the coarse label combined with reformulation consistently dominates. The multi-panel layout illustrates the dampening effect of dietary compensation; at 30%, the magnitude of both benefits from the coarse label and harms from the detailed label are maximized. Variation from Model A to F results in a structural bias of 18.9 kcal in Model E and 14.9 kcal in Model F for the coarse versus detailed contrast, quantifying the shift induced by incremental covariate adjustment.

Furthermore, the horizontal comparison from the demographic-only Model A to the cognitively adjusted models provides visual confirmation of the licensing effect. In Model A, trajectories for all four labeling strategies remain relatively clustered near the status quo baseline. However, as transportability frameworks incorporate socioeconomic status in Model D and the working memory deficit in Model E and F, trajectories fan out dramatically. Detailed label scenarios sharply detach from the baseline, plunging into severe QALY deficits and escalating NHS expenditures. Even under high compensation assumptions, the cognitively adjusted models demonstrate that the detailed label still crosses into net health harm. These divergences reinforce the conclusion that optimizing public health outcomes requires simple heuristic consumer signaling paired with mandatory structural supply-side interventions. Incident obesity cases in Model F correlate with a clinical cascade involving 135,781 cardiovascular disease cases over the decade.

Table 5.3. Health gains and costs of coarse and detailed consumer FOP labelling policies under 50% dietary compensation Policy counterfactuals using transported Population Average Treatment Effects (PATE).

Model	Year	A. Coarse Label					B. Detailed Label				
		Kcal	Δ BMI	Obesity Prev.	QALYs	NHS (£m)	Kcal	Δ BMI	Obesity Prev.	QALYs	NHS (£m)
Baseline: Status Quo (Secular Trend)											
	Y1	-	0.01	-	-	£11.5	-	0.01	-	-	£11.5
	Y5	-	0.06	-	-	£136.0	-	0.06	-	-	£136.0
	Y10	-	0.11	-	-	£325.0	-	0.11	-	-	£325.0
Model A	Y1	-2.3	-0.02	58.8k	2.1k	-£11.0	1.7	0.02	-41.3k	-1.5k	£8.1
(Demographics)	*[95% UI]*		[-0.02,-0.02]	[24.3k,107.6k]	[1.4k,2.8k]	[-£13.9,-£8.2]		[0.02,0.02]	[-81.8k,-11.9k]	[-2.1k,-1.0k]	[£6.2,£10.1]
	Y5	-2.3	-0.03	84.3k	22.0k	-£104.9	1.7	0.02	-61.2k	-15.8k	£77.1
	[95% UI]		[-0.03,-0.03]	[41.3k,139.5k]	[12.0k,47.9k]	[-£133.0,-£76.6]		[0.02,0.02]	[-109.6k,-24.7k]	[-36.8k,-7.5k]	[£56.6,£97.4]
	Y10	-2.3	-0.03	70.9k	50.0k	-£212.2	1.7	0.02	-50.3k	-35.7k	£156.3
	[95% UI]		[-0.03,-0.03]	[27.4k,124.9k]	[15.5k,118.9k]	[-£275.0,-£153.3]		[0.02,0.03]	[-104.5k,-14.8k]	[-95.6k,-3.9k]	[£113.4,£200.0]
Model B	Y1	-2.4	-0.02	60.6k	2.2k	-£11.7	2.3	0.02	-60.0k	-2.1k	£11.2
(+ Children)	*[95% UI]*		[-0.02,-0.02]	[24.7k,111.1k]	[1.5k,3.1k]	[-£14.5,-£8.7]		[0.02,0.02]	[-106.4k,-22.6k]	[-2.9k,-1.4k]	[£8.4,£13.9]
	Y5	-2.4	-0.04	89.3k	23.4k	-£110.8	2.3	0.03	-87.1k	-21.9k	£106.1
	[95% UI]		[-0.04,-0.03]	[42.8k,144.4k]	[13.6k,54.0k]	[-£139.5,-£79.2]		[0.03,0.03]	[-140.5k,-41.9k]	[-51.2k,-8.1k]	[£77.3,£133.3]
	Y10	-2.4	-0.04	74.8k	54.2k	-£225.7	2.3	0.03	-71.7k	-51.5k	£214.2
	[95% UI]		[-0.04,-0.03]	[30.8k,130.2k]	[20.2k,127.8k]	[-£285.3,-£160.4]		[0.03,0.03]	[-124.3k,-31.0k]	[-126.5k,-11.2k]	[£154.4,£273.7]
Model C	Y1	2.4	0.02	-62.8k	-2.2k	£11.7	12.9	0.12	-337.6k	-11.9k	£62.4
(+ Health)	*[95% UI]*		[0.02,0.02]	[-111.6k,-25.5k]	[-3.0k,-1.4k]	[£8.8,£14.5]		[0.12,0.12]	[-437.2k,-243.0k]	[-15.8k,-8.3k]	[£47.8,£77.2]
	Y5	2.4	0.04	-88.6k	-22.7k	£111.2	12.9	0.19	-490.2k	-134.1k	£591.1
	[95% UI]		[0.03,0.04]	[-147.1k,-43.4k]	[-53.5k,-8.3k]	[£83.6,£140.4]		[0.18,0.19]	[-616.3k,-366.2k]	[-196.0k,-86.6k]	[£453.6,£732.6]
	Y10	2.4	0.04	-75.8k	-52.7k	£225.1	12.9	0.19	-408.2k	-317.6k	£1197.3
	[95% UI]		[0.03,0.04]	[-125.2k,-30.1k]	[-126.5k,-13.0k]	[£163.2,£287.9]		[0.18,0.19]	[-527.6k,-302.2k]	[-458.9k,-192.3k]	[£902.7,£1496.6]
Model D	Y1	-0.5	-0.00	11.8k	483	-£2.6	12.5	0.12	-328.6k	-11.4k	£60.1
(+ SES)	*[95% UI]*		[-0.00,-0.00]	[0,38.0k]	[327,663]	[-£3.2,-£2.0]		[0.11,0.12]	[-428.2k,-232.1k]	[-15.2k,-7.9k]	[£46.2,£74.9]
	Y5	-0.5	-0.01	18.4k	4.8k	-£24.6	12.5	0.18	-472.9k	-128.7k	£570.6
	[95% UI]		[-0.01,-0.01]	[695,50.2k]	[3.0k,16.7k]	[-£31.0,-£17.4]		[0.18,0.18]	[-591.2k,-360.0k]	[-189.6k,-80.8k]	[£438.5,£712.7]
	Y10	-0.5	-0.01	16.1k	10.2k	-£50.2	12.5	0.18	-392.9k	-306.4k	£1153.6
	[95% UI]		[-0.01,-0.01]	[0,43.8k]	[4.1k,46.1k]	[-£63.6,-£29.9]		[0.17,0.18]	[-510.0k,-290.3k]	[-463.4k,-185.3k]	[£887.5,£1437.5]
Model E	Y1	-3.4	-0.03	86.2k	3.2k	-£16.5	10.0	0.09	-262.1k	-9.1k	£48.1
(+ $d'2$)	*[95% UI]*		[-0.03,-0.03]	[45.9k,141.8k]	[2.1k,4.3k]	[-£20.3,-£12.3]		[0.09,0.09]	[-354.8k,-181.1k]	[-12.1k,-6.2k]	[£36.6,£58.4]
	Y5	-3.4	-0.05	125.7k	33.7k	-£155.8	10.0	0.15	-371.3k	-101.5k	£455.1
	[95% UI]		[-0.05,-0.05]	[71.7k,188.7k]	[17.8k,68.9k]	[-£196.2,-£116.1]		[0.14,0.15]	[-486.8k,-272.9k]	[-154.1k,-63.0k]	[£347.7,£560.4]
	Y10	-3.4	-0.05	105.9k	79.3k	-£316.5	10.0	0.14	-316.3k	-237.1k	£922.6
	[95% UI]		[-0.05,-0.05]	[56.1k,172.3k]	[35.6k,166.4k]	[-£398.7,-£231.1]		[0.14,0.15]	[-420.0k,-222.9k]	[-372.9k,-144.2k]	[£701.5,£1141.0]
Model F	**Y1**	**-2.1**	**-0.02**	**53.0k**	**1.9k**	**-£10.2**	**9.3**	**0.09**	**-244.3k**	**-8.5k**	**£44.6**
(+ $d'3$)	***[95% UI]***		**[-0.02,-0.02]**	**[19.8k,101.0k]**	**[1.3k,2.7k]**	**[-£12.7,-£7.7]**		**[0.08,0.09]**	**[-326.1k,-167.8k]**	**[-11.4k,-5.6k]**	**[£32.7,£55.8]**
	Y5	**-2.1**	**-0.03**	**78.1k**	**19.9k**	**-£97.1**	**9.3**	**0.14**	**-350.4k**	**-94.8k**	**£423.5**
	[95% UI]		**[-0.03,-0.03]**	**[34.7k,131.8k]**	**[10.8k,47.7k]**	**[-£122.3,-£71.4]**		**[0.13,0.14]**	**[-452.1k,-253.6k]**	**[-139.3k,-58.4k]**	**[£314.2,£531.4]**
	Y10	**-2.1**	**-0.03**	**64.5k**	**45.7k**	**-£196.8**	**9.3**	**0.13**	**-294.0k**	**-222.0k**	**£856.3**
	[95% UI]		**[-0.03,-0.03]**	**[24.5k,119.5k]**	**[9.4k,111.3k]**	**[-£250.6,-£142.1]**		**[0.13,0.14]**	**[-395.3k,-201.3k]**	**[-349.8k,-123.8k]**	**[£638.0,£1082.7]**

Note: Point estimates represent the median cumulative impact over 1-year (Y1), 5-year (Y5), and 10-year (Y10) horizons for the simulated UK adult population ($N \approx 53$ million). 95% Uncertainty Intervals [UI] were derived from 1,000 Monte Carlo iterations of a Probabilistic Sensitivity Analysis (PSA). **Baseline (Status Quo)** represents the absolute burden of un-intervened natural weight gain (secular trend). All intervention scenarios represent the incremental difference compared to this baseline. **Avg. Daily Kcal Change** reflects the effective individual caloric deficit after enforcing a strict 50% dietary compensation rate ($\rho = 0.5$) for subsequent daily eating. Negative Incremental BMI values represent net weight loss (health benefit); negative NHS Costs represent net financial savings to the health system. Projections utilize causal effect estimates transported via TMLE-BART showed in Table 3. Abbreviations: k = thousands; m = millions.

Table 5.4. Health gains and costs of coarse and detailed FOP labelling combined with industry reformulation under 50% dietary compensation: Policy counterfactuals using transported Population Average Treatment Effects (PATE).

Model	Year	C. Coarse + Reformulation					D. Detailed + Reformulation				
		Kcal	Δ BMI	Obesity Prev.	QALYs	NHS (£m)	Kcal	Δ BMI	Obesity Prev.	QALYs	NHS (£m)
Baseline: Status Quo (Secular Trend)											
	Y1	-	0.01	-	-	£11.5	-	0.01	-	-	£11.5
	Y5	-	0.06	-	-	£136.0	-	0.06	-	-	£136.0
	Y10	-	0.11	-	-	£325.0	-	0.11	-	-	£325.0
Model A (Demographics)	Y1	-3.5	-0.03	92.5k	3.2k	-£17.0	0.4	0.00	-9.7k	-398	£2.1
	[95% UI]		[-0.03,-0.03]	[45.0k,150.1k]	[2.1k,4.4k]	[-£21.0,-£12.7]		[0.00,0.00]	[-33.9k,0]	[-559,-272]	[£1.6,£2.7]
	Y5	-6.0	-0.11	291.7k	54.8k	-£253.6	-2.1	-0.06	142.9k	14.5k	-£73.7
	[95% UI]		[-0.11,-0.11]	[210.2k,388.2k]	[30.8k,96.4k]	[-£315.7,-£190.9]		[-0.06,-0.05]	[88.0k,217.3k]	[8.3k,28.9k]	[-£92.7,-£53.0]
	Y10	-9.2	-0.20	443.2k	191.9k	-£779.7	-5.2	-0.15	317.4k	95.0k	-£417.3
	[95% UI]		[-0.21,-0.20]	[335.3k,566.4k]	[102.0k,302.4k]	[-£975.0,-£578.2]		[-0.15,-0.14]	[230.4k,420.1k]	[48.2k,164.6k]	[-£524.5,-£307.0]
Model B (+ Children)	Y1	-3.7	-0.03	93.4k	3.4k	-£17.6	1.1	0.01	-26.2k	-971	£5.1
	[95% UI]		[-0.03,-0.03]	[45.5k,152.1k]	[2.2k,4.5k]	[-£22.0,-£13.1]		[0.01,0.01]	[-58.9k,-4.9k]	[-1.3k,-623]	[£3.9,£6.4]
	Y5	-6.2	-0.12	294.5k	57.0k	-£258.4	-1.4	-0.05	119.8k	8.6k	-£44.1
	[95% UI]		[-0.12,-0.11]	[211.6k,399.6k]	[33.1k,91.7k]	[-£326.9,-£194.3]		[-0.05,-0.05]	[65.8k,190.3k]	[-2.9k,20.7k]	[-£56.2,-£33.1]
	Y10	-9.3	-0.21	447.2k	197.2k	-£785.2	-4.6	-0.14	299.7k	81.2k	-£352.8
	[95% UI]		[-0.21,-0.20]	[337.5k,570.6k]	[114.7k,301.7k]	[-£994.6,-£585.6]		[-0.14,-0.13]	[208.7k,410.7k]	[34.1k,135.9k]	[-£444.4,-£268.9]
Model C (+ Health)	Y1	1.2	0.01	-29.2k	-1.1k	£5.6	11.7	0.11	-305.7k	-10.7k	£56.1
	[95% UI]		[0.01,0.01]	[-65.4k,-6.9k]	[-1.5k,-707]	[£4.2,£7.0]		[0.11,0.11]	[-406.4k,-214.9k]	[-14.1k,-7.2k]	[£42.2,£70.5]
	Y5	-1.3	-0.05	115.5k	7.6k	-£39.8	9.2	0.11	-279.3k	-99.9k	£440.2
	[95% UI]		[-0.05,-0.04]	[66.5k,184.7k]	[-6.1k,17.8k]	[-£50.9,-£29.3]		[0.10,0.11]	[-383.3k,-187.9k]	[-153.5k,-60.2k]	[£329.7,£550.9]
	Y10	-4.5	-0.14	294.4k	76.0k	-£347.7	6.1	0.02	-32.2k	-173.0k	£629.0
	[95% UI]		[-0.14,-0.13]	[209.5k,392.9k]	[32.3k,138.6k]	[-£433.1,-£260.1]		[0.01,0.02]	[-72.5k,-1.8k]	[-306.8k,-84.0k]	[£458.9,£791.5]
Model D (+ SES)	Y1	-1.8	-0.02	43.9k	1.6k	-£8.6	11.2	0.10	-294.6k	-10.3k	£54.3
	[95% UI]		[-0.02,-0.02]	[14.1k,86.2k]	[1.1k,2.2k]	[-£10.8,-£6.3]		[0.10,0.10]	[-388.4k,-206.6k]	[-13.6k,-6.7k]	[£41.3,£67.4]
	Y5	-4.3	-0.09	221.4k	36.4k	-£175.2	8.7	0.10	-259.6k	-95.0k	£421.2
	[95% UI]		[-0.09,-0.09]	[151.4k,311.1k]	[20.3k,62.2k]	[-£220.4,-£129.3]		[0.10,0.10]	[-347.3k,-184.5k]	[-149.8k,-54.8k]	[£318.3,£524.8]
	Y10	-7.4	-0.18	388.6k	147.5k	-£621.4	5.6	0.01	-16.7k	-163.7k	£587.2
	[95% UI]		[-0.18,-0.17]	[289.4k,501.6k]	[79.6k,233.6k]	[-£779.8,-£459.6]		[0.01,0.01]	[-50.0k,9.6k]	[-291.0k,-74.6k]	[£436.3,£743.4]
Model E (+ $d'2$)	Y1	-4.7	-0.04	119.6k	4.3k	-£22.5	8.7	0.08	-228.5k	-8.0k	£42.0
	[95% UI]		[-0.04,-0.04]	[67.0k,183.1k]	[2.8k,5.7k]	[-£27.9,-£16.7]		[0.08,0.08]	[-320.5k,-151.3k]	[-10.8k,-5.6k]	[£31.7,£52.7]
	Y5	-7.2	-0.13	332.4k	66.6k	-£307.3	6.2	0.06	-159.9k	-70.4k	£305.0
	[95% UI]		[-0.13,-0.13]	[243.0k,431.2k]	[38.7k,105.9k]	[-£382.3,-£225.4]		[0.06,0.07]	[-241.7k,-96.4k]	[-112.9k,-39.9k]	[£231.5,£387.2]
	Y10	-10.3	-0.22	473.6k	219.1k	-£885.3	3.1	-0.03	59.1k	-101.7k	£353.8
	[95% UI]		[-0.22,-0.21]	[356.5k,610.6k]	[127.8k,338.0k]	[-£1106.8,-£650.1]		[-0.03,-0.02]	[24.6k,107.6k]	[-206.7k,-36.7k]	[£260.3,£455.2]
Model F (+ $d'3$)	**Y1**	**-3.4**	**-0.03**	**86.1k**	**3.1k**	**-£16.3**	**8.0**	**0.07**	**-212.0k**	**-7.3k**	**£38.6**
	[95% UI]		**[-0.03,-0.03]**	**[42.6k,140.6k]**	**[2.0k,4.2k]**	**[-£20.0,-£12.2]**		**[0.07,0.07]**	**[-285.1k,-143.7k]**	**[-9.8k,-5.0k]**	**[£28.7,£47.9]**
	Y5	**-5.9**	**-0.11**	**283.6k**	**52.7k**	**-£247.6**	**5.5**	**0.05**	**-138.1k**	**-61.7k**	**£274.5**
	[95% UI]		**[-0.11,-0.11]**	**[202.5k,381.5k]**	**[30.6k,85.9k]**	**[-£305.7,-£184.3]**		**[0.05,0.05]**	**[-207.0k,-81.0k]**	**[-104.7k,-34.1k]**	**[£205.1,£344.1]**
	Y10	**-9.0**	**-0.20**	**439.9k**	**186.4k**	**-£765.1**	**2.4**	**-0.04**	**79.4k**	**-82.5k**	**£291.2**
	[95% UI]		**[-0.20,-0.20]**	**[328.5k,564.8k]**	**[105.7k,287.8k]**	**[-£953.2,-£571.6]**		**[-0.04,-0.03]**	**[37.2k,137.2k]**	**[-182.0k,-27.0k]**	**[£207.4,£372.4]**

Note: Estimates include a staged industry reformulation effect capping at 5% caloric reduction by year 5. Point estimates represent the median cumulative impact over 1-year (Y1), 5-year (Y5), and 10-year (Y10) horizons for the simulated UK adult population ($N \approx 53$ million). 95% Uncertainty Intervals [UI] were derived from 1,000 Monte Carlo iterations of a Probabilistic Sensitivity Analysis (PSA). **Baseline (Status Quo)** represents the absolute burden of un-intervened natural weight gain (secular trend). All intervention scenarios represent the incremental difference compared to this baseline. **Avg. Daily Kcal Change** reflects the effective individual caloric deficit after enforcing a strict 50% dietary compensation rate ($\rho = 0.5$) for subsequent daily eating. Negative Incremental BMI values represent net weight loss (health benefit); negative NHS Costs represent net financial savings to the health system. Projections utilize causal effect estimates transported via TMLE-BART showed in Table 3. Abbreviations: k = thousands; m = millions.

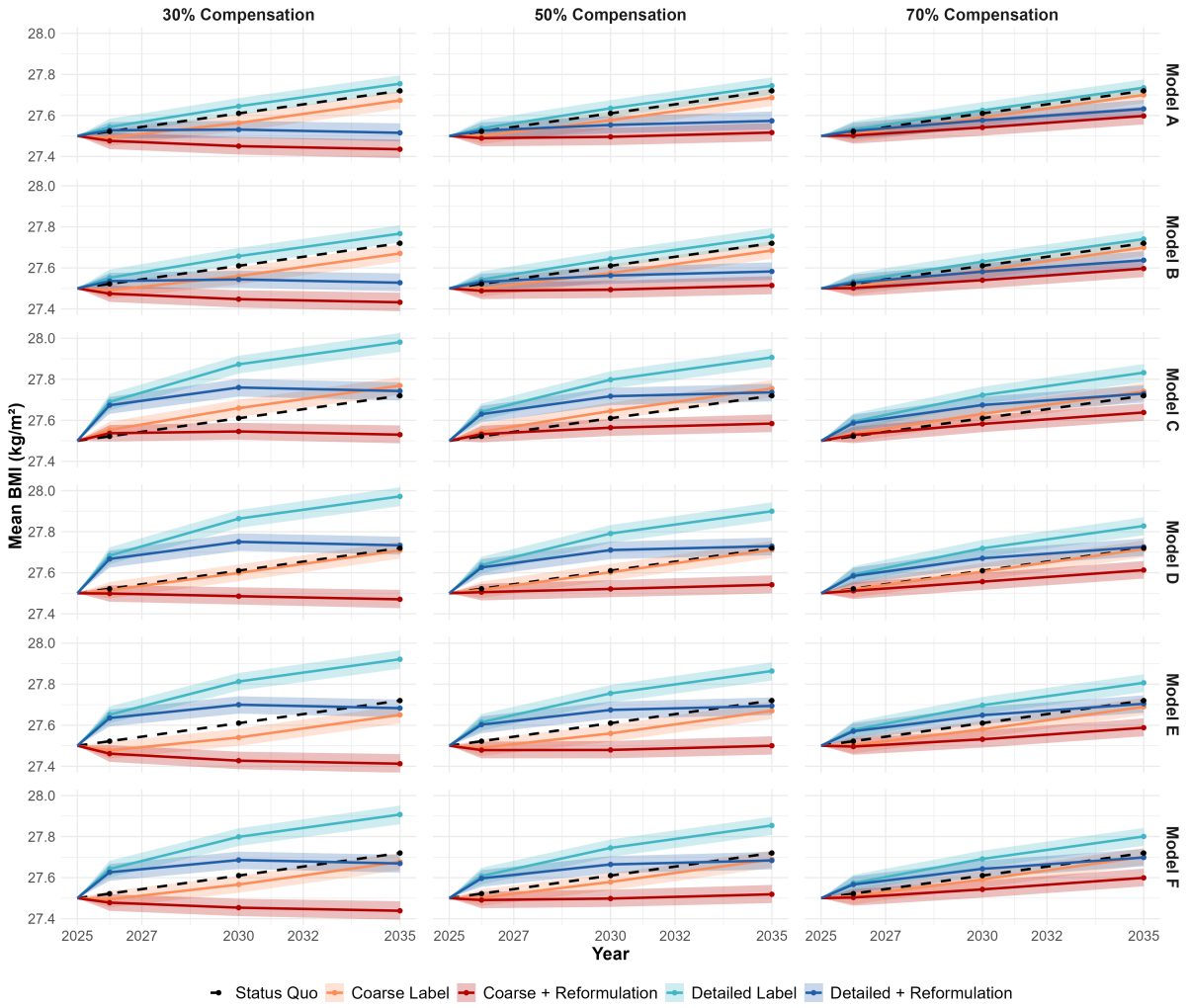

Fig. 5.4. Projected population mean BMI change over a 10-year horizon (2025—2035) under alternative FOP labelling policy scenarios. The Status Quo (Black line) represents the counterfactual baseline of natural weight gain (secular trend) in the absence of policy intervention. Colored lines represent the mean projected trajectories for the Coarse Label (yellow), Coarse Label + Industry Reformulation (red), Detailed Label (light blue), and Detailed Label + Industry Reformulation (Blue) scenarios, based on Population Average Treatment Effect (PATE) estimates from Model A to F. Estimates under 0.3%, 0.5% and 0.7% dietary compensation assumptions. The projection incorporates a stochastic noise parameter to simulate natural year-to-year population variability and an implementation lag function assuming partial policy effect in years 1—2 before reaching steady state. Shaded ribbons indicate the 95% confidence intervals derived from the uncertainty in the transported treatment effect estimates.

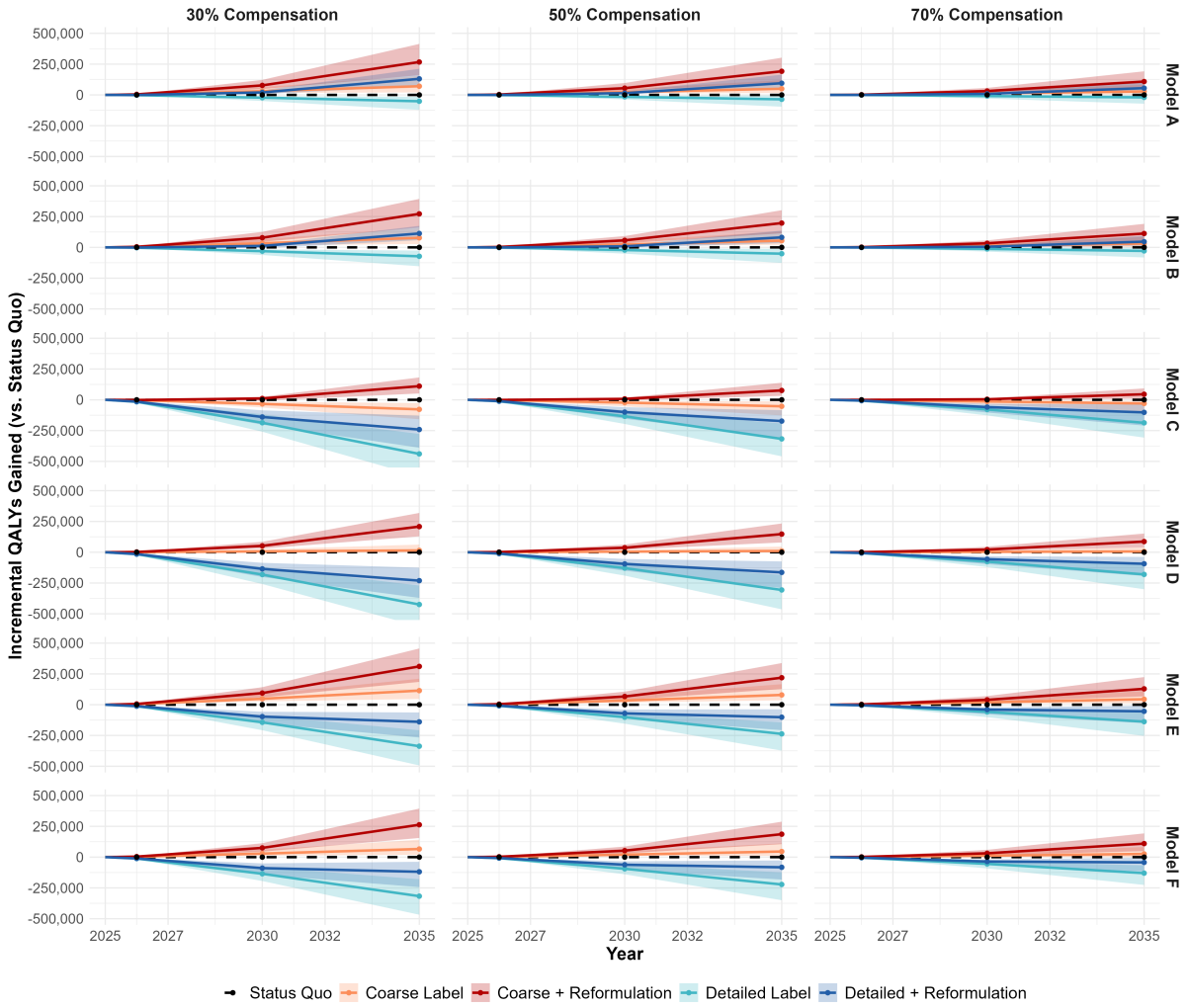

Fig. 5.5. Projected population mean NHS costs over a 10-year horizon (2025—2035) under alternative FOP labelling policy scenarios. The Status Quo (Black line) represents the counterfactual baseline of natural weight gain (secular trend) in the absence of policy intervention. Colored lines represent the mean projected trajectories for the Coarse Label (yellow), Coarse Label + Industry Reformulation (red), Detailed Label (light blue), and Detailed Label + Industry Reformulation (Blue) scenarios, based on Population Average Treatment Effect (PATE) estimates from Model A to F. Estimates under 0.3%, 0.5% and 0.7% dietary compensation assumptions. The projection incorporates a stochastic noise parameter to simulate natural year-to-year population variability and an implementation lag function assuming partial policy effect in years 1—2 before reaching steady state. Shaded ribbons indicate the 95% confidence intervals derived from the uncertainty in the transported treatment effect estimates.

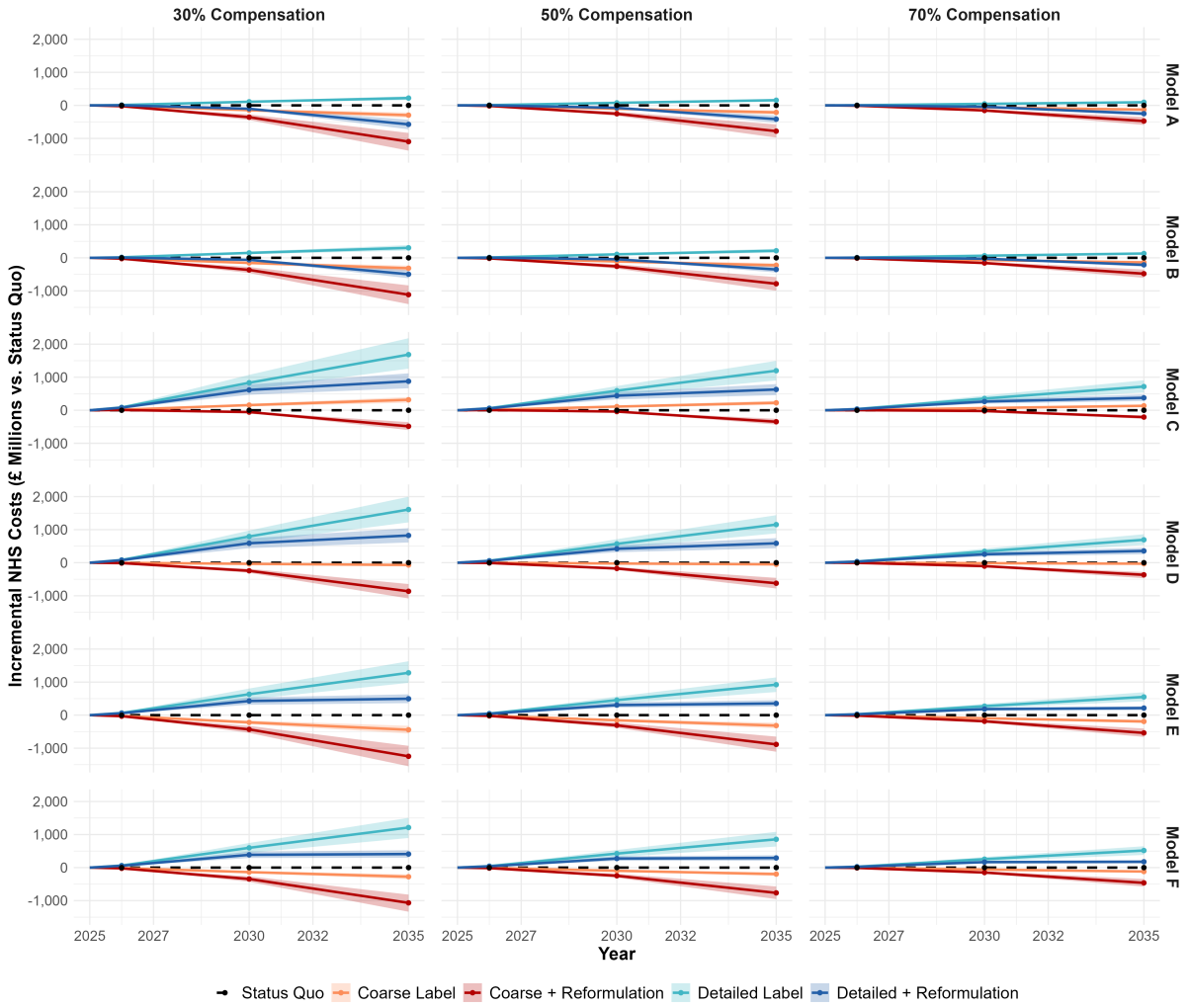

Fig. 5.6. Projected population mean QALY gained over a 10-year horizon (2025—2035) under alternative FOP labelling policy scenarios. The Status Quo (Black line) represents the counterfactual baseline of natural weight gain (secular trend) in the absence of policy intervention. Colored lines represent the mean projected trajectories for the Coarse Label (yellow), Coarse Label + Industry Reformulation (red), Detailed Label (light blue), and Detailed Label + Industry Reformulation (Blue) scenarios, based on Population Average Treatment Effect (PATE) estimates from Model A to F. Estimates under 0.3%, 0.5% and 0.7% dietary compensation assumptions. The projection incorporates a stochastic noise parameter to simulate natural year-to-year population variability and an implementation lag function assuming partial policy effect in years 1—2 before reaching steady state. Shaded ribbons indicate the 95% confidence intervals derived from the uncertainty in the transported treatment effect estimates.

5.4 Sensitivity analysis

To ensure our conclusions were not contingent on a single modeling approach, we conducted a robustness check by replicating the transportability analysis using a Generalized Linear Model-based Inverse Probability of Sampling Weighting (GLM-IPSW). The GLM-IPSW method estimates the probability of trial participation using logistic regression (a GLM) instead of the non-parametric Bayesian Additive Regression Trees (BART) used in the main analysis. This shift confirms that our findings are not solely dependent on the performance of a single machine learning technique. The results of the GLM-IPSW analysis consistently confirmed the primary findings, demonstrating robustness against model specification (see Appendix). For the primary specification (Model D), the GLM-IPSW PATE for the Detailed vs Control comparison was 19.6 kcal [95% CI: 3.3, 41.7], a statistically significant increase that corroborates the backfire effect identified in the main analysis. Similarly, the Coarse vs Detailed comparison in Model D showed a point estimate of -7.7 kcal [95% CI: -28.8, 9.7], indicating a directional benefit of the simpler label, although the confidence intervals for the Coarse label were wider in the GLM specification. Crucially, the concordance across two fundamentally different estimators lends credibility to the stability and validity of the core estimates.

A further robustness check was performed by stratifying the target population into distinct temporal eras using the DR-BART model to assess the stability of the PATEs over time. We examined the stability of the relative benefit of the Coarse label over the Detailed label (Coarse vs Detailed PATE) across the post-MTL period (2013—2018) and the COVID/post-COVID period (2019—2022). As shown in the Appendix, the PATE estimates for Model D remained directionally consistent, with the Coarse label consistently outperforming the Detailed label (indicated by negative PATE values). In the early period (2008—2013), this benefit was statistically significant in several years, peaking at -23.2 kcal in 2010. While the magnitude of the effect fluctuated during the mid-period (2014—2018), the direction of effect remained stable. Notably, during the pandemic period (2019—2022),

the advantage of the Coarse label re-emerged strongly, with significant PATEs of -17.0 kcal [95% CI: -37.5, -1.4] in 2021 and -14.4 kcal [95% CI: -33.3, -1.5] in 2022. This temporal consistency confirms that the relative superiority of the simpler Coarse label is stable and not driven by short-term fluctuations in consumer behavior or dataset composition.

Finally, given the critical role of working memory (d') as a moderator in our model, we conducted an operational validation of the synthetic cognitive data generation process (see Appendix). We compared two generation methods: an XGBoost model with monotonic age constraints (Panel A) and a Calibrated Linear Regression model with a fixed biological slope of -0.012 points/year (Panel B). While the unconstrained machine learning approach captured local variances, the linear model enforced a strict biological gradient. However, due to the overly rigid constraints of the linear specification which failed to capture non-linear distributional characteristics of the population data, the XGBoost approach was retained for the primary analysis.

As a form of operational validation, the robustness of the microsimulation core findings was subjected to cross-validation testing. The primary health outcome BMI over 10 years and economic outcome Cost/QALY were compared against published cost-effectiveness literature concerning dietary interventions and food policy. For instance, the projected 0.85 kg/m2 BMI impact and the dominant status of the Detailed Label Reformulation policy were benchmarked against similar models for salt and sugar reduction. This included comparing the cost-effectiveness against salt reformulation interventions which were estimated to be dominant over a 10-year period using the PRIMEtime CE model, and analyses of salt reduction policies which found high cost-effectiveness for population-level interventions. Furthermore, the results align with modeling efforts for other dietary guidelines, such as incorporating new sugar and fibre guidelines, which also demonstrated favorable health and economic implications. The cross-validation confirmed that our model estimates for cost-effectiveness were broadly consistent with the direction and typical thresholds observed in the field, reinforcing the policy conclusion that mandated reformulation is the key driver of societal benefit.

5.5 Discussion

This study aimed to estimate the PATE of FOP nutritional labelling by assessing the methodological implications of omitting working memoryâ€"a theoretically grounded effect modifierâ€"in transportability frameworks. Drawing on data from a recent breakfast cereal choice trial examining label granularity [148], with the NDNS as the target population dataset, we found that overlooking discrepancies between volunteer trial participants and the UK target population leads to misestimated policy inferences [32, 43, 158, 150]. Trial participants exhibited systematically higher working memory capacities (mean standardized 2-back d' score of 3.21 vs. 2.74 in the target population), substantially greater educational attainment (86.6% holding college degrees vs. 67.4% nationally), and marked overrepresentation of individuals actively attempting weight loss. These stark disparities relative to the target population underscore the need for advanced causal adjustment techniques [136, 138, 145, 135].

Our research empirically validates the theoretical warnings posited in recent causal inference literature regarding unobserved moderators [31, 44, 41, 140]. As hypothesized, omitting working memory severely biased the PATE. By evaluating hierarchical nested models, we demonstrated that relying solely on standard demographic and socioeconomic proxies failed to capture the true treatment heterogeneity [141, 142]. To illustrate this, we focused on the treatment comparison between the coarse and detailed labelsâ€"the contrast that yielded the most robust statistical significance across our fully specified modelsâ€"â€"in breakfast cereal choices. Transporting this causal effect using only standard demographic proxies yielded a misleading null finding, estimating an average reduction of -8.0 calories [95% CI: -22.7, 7.1]) in favor of the coarse label. This comparison closely mirrors the unadjusted, naive trial estimates, which indicated a non-significant relative effect of -5.7 mean calories [95% CI: -19.0, 7.6].

Adjusted models incorporating synthetically generated d' scores within the benchmarked transportability framework revealed significant relative reductions in caloric counts. Model

E and F showed reductions in average calories of -26.9 [95% CI: -47.5, -5.8] and -22.9 [95% CI: -41.3, -5.2], respectively, in favour of the coarse label. The explicit 18.9 average calorie shift between the basic demographic adjustment of Model A and the fully cognitively adjusted Model E directly quantifies the structural bias induced by unmeasured confounding. This analytical difference aligns with the literature advocating accommodation of the variance penalty inherent in generative synthetic dataâ€"a statistically superior strategy to ignoring unmeasured heterogeneity that enables policymakers to reliably estimate the true PATE [140, 142, 44].

To rigorously defend the use of generative synthetic data for imputing a missing confounder against statistical uncertainty, we evaluated PATE estimates under extreme scenario bounds [143, 54, 142]. We artificially perturbed the synthetically generated d' scores by -10% and $+10\%$ to establish worst-case parameters for the unobserved confounder. For Model E, the coarse label's relative superiority over the detailed labelling proved highly robust. Under the lower-bound scenario, this advantage intensified, yielding a relative reduction of -31.2 calories [95% CI: -52.3, -9.8] on average versus the detailed label. Conversely, in the upper-bound scenario—simulating an unexpectedly capable population—the coarse label still maintained a statistically significant relative reduction of -24.4 calories [95% CI: -46.0, -3.4] on average. Because the coarse label's directional superiority remains structurally stable and statistically significant across these deterministic extremes, we can confidently conclude that the exact precision of the generative imputation does not threaten the fundamental validity of the policy evaluation [143].

We interpret this stark reversal in the causal effect from the trial estimate to the PATE through the lens of cognitive load theory [27]. This interpretation challenges the classical assumption in public health that merely providing consumers with more detailed information automatically nudge healthier behaviors [12, 19, 173, 17]. Instead, we posit that the fundamental prerequisite for FOP label effectiveness is consumers' ability to accurately process and understand the provided message while accounting for preference biases [174, 36, 21]. The original trial explicitly tested this by manipulating the number

of information chunks, confirming that label effectiveness is strictly bounded by consumer working memory [148]. Specifically, the trial interaction models demonstrated that each one unit increase in $d'2$ score reduced average calories selection by -15.2 calories [95% CI: -30.1, -0.3] under coarse labels and by -19.2 calories [95% CI: -33.8, -4.6] under detailed labels when compared to absent labels.

While higher d' scores enhance calorie reductions under both label formats, more granular labels prove optimal only for high-capacity consumers, such as the volunteer trial participants who were adept at parsing intricate nutritional information in a breakfast cereal selection task [148, 36]. We interpret this to mean that the detailed label may overwhelm the typical consumer in the cognitively heterogeneous target population [67, 130, 175]. For the average consumer, the intricate information in the eight-chunk label exceeds processing limits, causing severe cognitive overload [27]. As a result, rather than serving as an objective informational tool, the complex label devolves into a superficial health halo [146, 7]. For instance, consumers might perceive a high-sugar breakfast cereal displaying comprehensive nutrient details as overall healthy, justifying larger portions or added toppings despite its elevated caloric intake.

The overload activates moral licensing [147], where the mere presence of extensive data might grant consumers psychological permission to indulge, allowing them to rationalize choosing more energy-dense products without fully processing the actual caloric costs. Despite this theoretical importance, the current literature offers little guidance on the optimal number of information chunks to maximize usability across heterogeneous audiences, often leaving regulators to determine label granularity on pragmatic or political grounds rather than cognitive evidence [129].

This study demonstrates that transportability sensitivity analysis for missing confounders in PATE estimation provides essential safeguards for cost-effectiveness policy evaluations. Our 10-year dynamic microsimulation model, which incorporates the fixed PATE, revealed that biased proxy-based estimates distort projections by masking adverse behavioral responses, leading to substantial misestimations of noncommunicable disease prevalence

and healthcare costs in the UK [155, 152]. Specifically, under the detailed labelling counterfactual, health outcomes worsened progressively from Model A to Model F. Due to moral licensing-induced weight gain, Model A predicted 50.3k additional obesity cases by year 10, whereas coarse labelling prevented 70.9k cases assuming 50% dietary compensation.

Drastically, Model F predicted 294.0k additional cases under detailed labelling, while coarse labelling still prevented 64.5k. The discrepancy between Models A and F under the detailed labelling scenario underscores the extent of misestimation arising from the failure to account for working memory, the true confounder. This scale of obesity cases aligns closely with effect sizes reported in prior FOP labelling studies. Our microsimulation projected an upstream increase of 243,700 incident obesity cases. As demonstrated by [151], such population BMI shifts are critical, as they cascade into severe downstream clinical endpoints, such as the 135,781 cardiovascular disease cases seen in similar microsimulation models [153].

Microsimulation modeling confirms that maximal public health benefits are achieved through structural supply-side measures. The counterfactual combining coarse labeling with mandatory 5% industry reformulation emerged as the superior policy option. In contrast, pairing the detailed label with reformulation underscored the severe risks of unmeasured confounding. Model A, reliant on biased demographic proxies, portrayed this combination as highly effective, falsely projecting the prevention of 317,400 obesity cases, a gain of 95,000 QALYs, and £417.3 million in NHS cost savings by year 10. Conversely, the fully adjusted Model F revealed the opposite outcome: the complex label's behavioral rebound completely offset reformulation's structural benefits, resulting in 79,400 new obesity cases, a net loss of 82,500 QALYs, and £291.2 million in additional NHS expenditures over the decade. This stark reversal demonstrates that, although industry reformulation constitutes an essential precondition for labeling effectiveness [13, 19], it cannot mitigate the harmful behavioral consequences of cognitive overload induced by elaborate information.

The disparities among these longitudinal trajectories narrowed as dietary compensation intensified [20]. Given that individuals typically compensate for caloric perturbationsâ€"such as the cereal selections for breakfastâ€"via compensatory intake across subsequent meals, we modeled outcomes at 30%, 50%, and 70% compensation levels. At the most conservative 30% threshold, health benefits and harms attained their peak magnitudes, markedly amplifying the divergence between proxy-dependent Model A and cognition-adjusted Model F. As compensation escalated to 70%, the physiological effects of the initial labelling exposure attenuated markedly, contracting all policy outcomes relative to standard 50% benchmarks. Critically, this dampening notwithstanding, the underlying structural bias endured intact across sensitivities; even in the maximally diluted 70% case, Model F's detailed label trajectory veered sharply from Model A's sanguine forecasts into unequivocal net harm. This substantiates that unmeasured cognitive capacity-driven licensing surpasses conventional physiological attenuation, affirming the imperative of rigorous causal adjustment irrespective of posited adherence behaviors [154].

In summary, this study advances both causal inference methodology and public policy insights. Methodologically, it highlights the vital role of sensitivity analysis for omitted confounders in PATE estimation to enable causal generalization. It demonstrates that ensuring transportability validity requires deliberately reconstructing theoretical effect modifiers, confirming that addressing unobserved confounders demands moving beyond conventional theoretical constraints and routine demographic proxies. Substantively, the analysis provides a cautionary population-level assessment of nutritional labeling, revealing that excessively detailed informationâ€"without structural interventionsâ€"is counterproductive in a cognitively heterogeneous population. In conclusion, we propose that straightforward coarse labeling, coupled with mandatory industry reformulation, offers the optimal economically and clinically effective approach for the UK.

5.6 Limitations

The primary strength of this paper lies in its rigorous application and comparison of modern transportability methods to a critical public health question. By leveraging a doubly robust estimator with a flexible machine learning algorithm, conducting an extensive sensitivity analysis, and explicitly tackling the methodological challenge of covariate selection, our work provides a robust template for future research in this domain. However, this study is subject to several limitations that boundary the generalizability of our findings.

First, the core limitation stems from the external validity of the source trial. The study incorporated a single food category and was conducted entirely in an online setting, which offered fewer distractions and allowed for digital filtering compared to busy physical store environments. This online setting, combined with a highly literate and technology savvy volunteer sample, restricts the direct generalizability of our findings to real world shopping behavior. Factors such as brand loyalty, pricing, shelf placement, and the specific marketing cues present in physical stores were not considered but substantially influence consumer behavior. Consequently, the current findings remain strictly bounded by the United Kingdom context and the specific covariates measured.

Second, a substantial temporal disparity exists between the experimental trial data, collected in late 2024, and the observational target data pooled from 2008 to 2023. This time gap introduces the potential for temporal drift. Rapidly evolving dietary habits, shifting structural food environments, and potential generational changes in cognitive abilities across cohorts might affect the longitudinal stability of the transported estimates.

The synthetic generation of working memory scores for the target population is a critical methodological step that introduces inherent approximations. While generated through a principled data driven process, the cognitive variable remains a simulation and cannot fully substitute for genuine population level measurements. This synthesis process inevitably

introduces assumptions that impact the precision and uncertainty of the transported estimates, specifically widening the confidence intervals of the interaction effects. The necessary trade off to correct for selection bias results in variance inflation, meaning the exact precision of the cognitive component of the population average treatment effect must be interpreted cautiously.

While the doubly robust Bayesian additive regression tree approach is a powerful tool, it presents specific methodological boundaries. The entire causal framework relies on the strict assumption of conditional mean exchangeability. If unmeasured variables exist that jointly influence trial participation and calorie consumption, this assumption is violated, compromising the validity of the causal claims. Furthermore, while the algorithm organically captures complex non linear relationships, relying on it without deeper structural interpretation risks treating the estimation as an opaque algorithmic process, potentially obscuring the specific localized mechanisms of effect modification.

The ten year microsimulation model relies on several simplifying assumptions that limit its ecological validity. The uniform industry reformulation assumption of a five percent caloric reduction fails to capture the dynamic heterogeneity of real world industry responses, which vary widely across different food categories and manufacturer sizes. Additionally, applying a fixed dietary compensation factor of 0.5 oversimplifies compensatory eating, which is highly variable, context dependent, and influenced by individual metabolic characteristics. Finally, the simulated outcomes are sensitive to the assumed health utility decrements, the noise process parameters, and the long term sustainability of the programmed secular trend.

To address these limitations, future research should advance both substantive public health evaluations and causal methodology. Substantively, validation studies with diverse participant groups are required to test how different levels of information chunk optimization directly influence consumer understanding and healthier food choices across varying cognitive abilities, specifically measuring the practical implementation feasibility of simplified labels in physical grocery environments. Furthermore, future public health modeling

should incorporate more heterogeneous and realistic parameters for industry reformulation, varying reductions based on food category and manufacturer size or historical trends. Researchers should also explore the sensitivity of long term health projections to dynamic, individualized models of dietary compensation rather than fixed aggregate factors. The scope of future generalizability efforts must also be broadened to encompass different cultural contexts outside the United Kingdom, explicitly exploring the unmeasured impact of localized marketing strategies and individual dietary preferences.

Methodologically, future research must investigate alternative, more robust methods for imputing unobserved effect modifiers in target populations to reduce bias while actively mitigating the severe variance inflation associated with incorporating multiple synthetic covariates. Additionally, analytical frameworks must be developed to formally account for temporal covariate shift when transporting effects across misaligned time periods. Finally, further statistical work is needed to explore the sensitivity of these results to potential violations of conditional mean exchangeability, alongside developing new techniques to interpret tree based transportability models more deeply, allowing researchers to unpack the mechanisms of effect modification rather than treating the predictive engine as a black box.

In summary, this work advances both substantive policy and causal methodology. Substantively, it delivers a cautionary population level appraisal of nutritional labeling, demonstrating that highly detailed information without structural support is counterproductive for a cognitively diverse public. We conclude that simple coarse labeling paired with mandatory industry reformulation is the most economically and clinically viable strategy for the United Kingdom. Methodologically, we corroborate the indispensability of generative machine learning coupled with doubly robust estimators for causal generalization. We prove that analytical validity in transportability depends entirely on the deliberate reconstruction of theoretical effect modifiers, confirming that overcoming the unobserved confounder problem requires moving beyond theoretical bounds and standard demographic proxies.

5.7 Conclusion

This research empirically substantiates the bias-variance trade-off inherent in transportability analyses, revealing that neglecting potent non-demographic effect modifiers generates erroneous population-level inferences. Utilizing doubly robust Bayesian additive regression tree (BART) estimators, we extrapolated trial results to the UK National Diet and Nutrition Survey (NDNS) population, establishing that adjustment for working memory capacity was indispensable to redress healthy volunteer bias, notwithstanding the associated variance increment. Substantively, these transported estimates refute the trial-derived presumption of labelling efficacy. The pronounced divergence between the proxy-dependent Model A and the fully cognition-adjusted Model F illuminates the acute hazards of unmeasured confounding. Whereas Model A, predicated on demographic proxies, fallaciously prognosticated that detailed labels allied with industry reformulation would bolster population health, Model F disclosed a pronounced backfire effect. Detailed labels engendered surplus caloric intake in the broader populace, ascribable to cognitive overload and moral licensing, whereby consumers justify calorie dense cereal options. This recalibration precipitated a stark bifurcation in forecasted health outcomes across the naÃ¯ve and adjusted models, evincing that standalone informational provisions are insufficient and potentially deleterious. By contrast, coarser labels reliably forestalled such adverse reactions. Model F's comprehensive health-economic forecasts validate coarse labelling conjoined with mandatory industry reformulation as the pre-eminent strategy, projected to forestall 439,900 obesity cases, accrue 186,400 quality-adjusted life years (QALYs), and deliver £765.1 million in National Health Service (NHS) savings across a decade. Ergo, maximising public health gains imperatives simplified consumer-oriented signals fused with supply-side structural interventions.

Chapter 6

Discussion

6.1 Synthesis of empirical evidence

The collective evidence generated through this thesis demands a fundamental shift in how researchers and policymakers conceptualize the pathway from policy implementation to public health impact. Traditionally, front-of-package (FOP) labeling has been treated as a straightforward tool for informational disclosure, resting on the neoclassical assumption that providing consumers with comprehensive data will naturally lead to optimized dietary consumption [129, 98]. However, by synthesizing the findings from the empirical investigations conducted in this dissertation, a more complex and nuanced reality emerges. This thesis demonstrates that the attenuated effectiveness of current labeling policies is not a failure of consumer willpower, but rather a systemic oversight regarding the perceptual and cognitive constraints of human information processing [11, 112].

6.1.1 Cognitive ease, salience, and consumer purchasing behavior

Addressing the primary aims of evaluating how the cognitive ease of interpreting FOP labels influences purchasing behavior and investigating the mechanisms of label salience, this research highlights critical perceptual and cognitive bottlenecks. In Chapter 3, the application of the quasi-experimental non-equivalent dependent variable design allowed for the isolation of print size readability effects from underlying nutritional content [176, 70]. The interpretation of these results reveals a primary perceptual barrier that actively hinders the salience-to-understanding pathway. When guidelines permit a minimum font size of

1.2 mm [63], they inadvertently establish a physical threshold that prevents the cognitive processing of the label for less-engaged consumers [6]. The findings suggest that while the UK's Multiple Traffic Light (MTL) system was initially associated with reductions in the frequent consumption of unhealthy ready-to-eat meals, temporal habituation to visual cues and supply-side adjustments by firms have mitigated these gains [19, 13]. This dynamic demonstrates that information provision is not static; interventions that succeed initially may be rendered ineffective as consumers habituate to the choice architecture.

Furthermore, analyzing how heterogeneity in consumer bandwidth shapes dietary decisions, the cognitive stratification dynamic identified in Chapter 4 provides the mechanistic link in this behavioral debate. By manipulating label granularity in a randomized controlled trial, the results demonstrate that augmented informational detail is only beneficial for individuals possessing the available cognitive capacity to process it [177, 111]. The statistical interaction between working memory (n-back performance) and label effectiveness indicates that highly detailed, multi-nutrient labels impose a disproportionate cognitive burden on individuals with lower working memory capacity. If only cognitively advantaged consumers can successfully utilize complex labels to differentiate products, the broader community is left to rely on visual heuristics that are easily manipulated by packaging design [146, 173].

6.1.2 Modeling population-level consumption patterns

To fulfill the aim of modeling how specific behavioral shifts ultimately alter population-level consumption patterns, Chapter 5 directly addressed the conceptual gap between immediate purchasing behavior and long-term dietary intake. When evaluating interventions at the national scale using the National Diet and Nutrition Survey (NDNS) data, the idealized behavioral effects observed in experimental trials systematically attenuate [178, 32].

The empirical discovery of a counterproductive behavioral response—wherein detailed labels induce excess caloric intake among specific demographic subgroups—provides a

direct, quantitative critique of the health halo phenomenon [146, 147]. This physiological rebound highlights a crucial distinction established in this thesis: optimizing point-of-purchase behavior does not guarantee optimized long-term consumption, as choices are heavily filtered through compensatory eating patterns, moral licensing, and cognitive overload. The analysis confirms that omitting latent effect modifiers, such as working memory, in generalization models introduces a substantial upward bias, projecting artificial, population-wide health benefits and providing policymakers with an overly optimistic estimation of intervention success [31, 43].

6.2 Methodological implications: bridging experimental evidence and real-world impact

To bridge the persistent gap between experimental evidence and real-world impact, a central tenet of this thesis is that rigorous causal methodology is a strict prerequisite for sound policy evaluation. By formalizing an advanced statistical framework—a tripartite causal inference pipeline spanning observational diagnosis, experimental identification, and policy generalization—this work provides a highly adaptable and reproducible template for the analysis of public health interventions. Following the statistical standards established in contemporary causal inference literature [179], this thesis demonstrates that the integration of generative machine learning into transportability methods represents a robust standard for navigating non-nested designs and external validity threats.

A primary methodological contribution centers on the critical selection of covariates for transportability and the resulting mitigation of bias [44]. The analysis establishes that the validity of causal generalization cannot rely solely on the convenience of available sociodemographic variables. Instead, the inclusion of theoretically motivated effect modifiers—namely, working memory—is absolutely critical for the unbiased estimation of population average treatment effects (PATE) [53]. This bias, driven by concurrent

mechanisms of cognitive overload and moral licensing, demonstrates that transportability models are exceptionally sensitive to the structural selection of effect modifiers. By employing advanced algorithmic techniques (XGboost) to systematically account for missing cognitive covariates across datasets, the methodology proves that accepting the variance penalty inherent in generative imputation is statistically and practically superior to relying on flawed demographic proxies. Failing to measure and transport the true behavioral constraints of the target population does not merely dilute the estimated effect; it can fundamentally reverse the policy recommendation.

Furthermore, the integration of these transported estimates into the cost-effectiveness microsimulation establishes a critical methodological bridge between causal inference and health economics. It demonstrates how finite-sample statistical biases compound severely over longitudinal forecasting horizons. The simulation reveals that labeling functions strictly as a supplementary intervention; the counterfactual scenarios resulting in the most mathematically significant reductions in obesity-related quality-adjusted life year losses are those where labels serve as a regulatory precursor to mandatory industry reformulation [164, 171]. This paradigm shift necessitates reevaluating labeling efficacy not merely by individual consumer comprehension metrics, but by the capacity of the intervention to force systemic, supply-side product modifications [180, 181].

6.3 Limitations and future directions for causal generalizability

Despite the methodological rigor applied across these three empirical chapters, this thesis is subject to several limitations that boundary the generalizability of the findings and dictate essential directions for future research.

First, regarding the observational evaluation in Chapter 3, the primary limitation lies in the persistent threat of unobserved confounding inherent in quasi-experimental designs

[70]. While the non-equivalent dependent variable design successfully controlled for broad historical trends and maturation effects, it could not account for highly localized, product-specific confounders, such as fluctuations in brand-level advertising expenditure or targeted pricing promotions that may have coincided temporally with the MTL implementation. Furthermore, national surveys rely on self-reported dietary consumption, which is highly susceptible to social desirability and recall bias [182]. To overcome this limitation, future public health evaluations must transition toward high-frequency digital data. Utilizing real-time grocery transaction logs or digital loyalty card datasets will allow researchers to capture genuine decision pathways *in situ*, thereby minimizing recall bias.

Second, the findings of the randomized controlled trial in Chapter 4 are constrained by the ecological validity of the experimental environment. The study was conducted online, which inherently filters out the physical and cognitive distractions characteristic of a busy supermarket setting, including time pressure, shelf placement, and competing marketing cues. Additionally, the trial focused exclusively on a single food category. Consequently, the isolated cognitive interaction observed between working memory and label granularity requires broader ecological validation. Future research must measure the practical implementation feasibility of optimized information chunks in physical retail environments, potentially utilizing mobile eye-tracking technologies to objectively capture visual attention pathways.

Third, the transportability and microsimulation framework developed in Chapter 5 relies on several complex structural assumptions. The process of bridging the non-nested design between the experimental trial and the NDNS target population necessitated the generative imputation of working memory scores. While employing advanced algorithmic imputation is methodologically superior to variable omission, it introduces inherent approximations and a variance penalty that widens the confidence intervals of the transported interaction effects. Furthermore, the doubly robust estimator fundamentally relies on the strict, untestable assumption of conditional mean exchangeability [32, 92]. While this study successfully adjusted for cognitive capacity, other latent effect modifiers—such as digital

literacy, baseline nutritional anxiety, or cultural food preferences—remain unmeasured and could introduce residual bias into the transported PATE estimates. Additionally, there is a substantial temporal disparity between the trial data collected in 2024 and the national survey target data pooled from 2008 to 2023. This discrepancy introduces the risk of temporal covariate shift.

To address these limitations, future research must prioritize the empirical collection of cognitive and behavioral covariates directly within large-scale national health surveys, thereby bypassing the need for synthetic imputation and strengthening the ignorability assumption. Methodologically, future work should focus on developing advanced sensitivity analysis frameworks specifically tailored for non-parametric machine learning estimators, such as BART, to better bound the uncertainty introduced by unmeasured confounding in transportability contexts.

Finally, the dynamic 10-year microsimulation relies on specific simplifying assumptions, including fixed dietary compensation aggregates and uniform industry reformulation parameters. Compensatory eating is a highly individualized physiological response rather than a static population constant. Future health economic modeling must explore the sensitivity of long-term health projections to dynamic, individualized models of dietary compensation. Furthermore, investigating the longitudinal dynamics of these behavioral responses—specifically, how cognitive overload and moral licensing evolve as populations habituate to granular label designs over a decade—will be crucial for accurate, long-term public health modeling.

6.4 Concluding remarks: health inequalities and policy design

A central motivation for the implementation of FOP policies, particularly the UK's Multiple Traffic Light system, is their theoretical capacity to address pervasive social

inequalities in health. In principle, utilizing universally recognized color heuristics is an attempt to democratize nutritional information, ensuring it is accessible and actionable for all consumers regardless of their socioeconomic, educational, or health literacy background. However, the empirical evidence synthesized in this thesis reveals a critical paradox: the specific granular design of the MTL may inadvertently exacerbate the very disparities it was intended to resolve [13].

By requiring consumers to process multiple informational chunks and weigh complex nutrient trade-offs simultaneously across fats, sugars, and salts, the policy inadvertently caters to individuals with higher cognitive bandwidth and baseline health literacy. As demonstrated through the causal framework, less-engaged consumers—or those facing cognitive overload in high-pressure shopping environments—are significantly disadvantaged by detailed labels [27]. They remain susceptible to perceptual bottlenecks, the health halo effect, and counterproductive compensatory eating behaviors. Therefore, rather than leveling the playing field, highly granular labeling schemes can effectively subsidize the health of the cognitively advantaged, leaving vulnerable populations to navigate an increasingly complex choice architecture without adequate heuristic support [12].

This dissertation concludes that the current trajectory of highly detailed nutritional labeling policy is reaching a definitive cognitive upper bound. The limits of what isolated, granular informational provision can achieve in public health have been largely exhausted. To advance dietary health outcomes and genuinely tackle diet-related health inequalities, policymakers must recognize that optimal nutrition is fundamentally a structural responsibility rather than an individual mandate. The empirical evidence presented herein demonstrates that coarse labels—characterized by simple, evaluative, and highly salient cues—provide a far implementation strategy for a general population possessing diverse cognitive resources [177, 7].

Public health frameworks must move beyond the narrative of individual consumer accountability. Consumers frequently opt for high-calorie, ultra-processed meals because these products constitute the most salient, convenient, and cognitively undemanding options

in their proximate environment [57, 55]. An authentically effective and equitable public health strategy should harness the causal insights derived from this thesis to enforce comprehensive industry reformulation, while concurrently implementing nutritional labels as intuitive heuristics rather than intricate analytical exercises. The paramount aim is to foster a regulated food environment in which the healthiest choice imposes the minimal demand on working memory, thereby emerging as the most instinctive and immediate selection for all consumers.

Appendices

Chapter A

Appendix Chapter 3

A.1 Food Frequency Questionnaire and Study Measures

QUESTION: At the moment, how often do you eat INSERT FOOD?

Alternatives:

- Cuts or portions of beef, lamb or pork, for example joints, steak, chops
- Burgers
- Sausages
- Chicken or turkey
- Duck or goose
- Pre-cooked meats
- Cured or dried meats
- Milk and dairy foods like cheese and yoghurt (INTERVIER NOTE: THIS INCLUDES DRINKING MILK, MILK IN TEA ETC.)
- Raw milk – by raw milk I mean milk that has not been pasteurised.
- Cooked eggs

- Cooked eggs? Please also think about food containing cooked eggs.
- Raw or uncooked eggs, including in things like homemade mayonnaise or homemade royal icing
- Cooked or smoked fish, excluding shellfish
- Cooked shellfish, for example crab, prawns, lobster, mussels
- Raw oysters
- Raw fruit
- Raw vegetables, including salad
- Cooked vegetables
- Frozen fruits, for example berries
- Pre-packed sandwiches
- Ready meals

Table A.1. Brant Test of the Parallel Regression Assumption for the ordinal logistic regression models. Tests were conducted on the unweighted complete-case dataset ($N = 9207$).

Variable	Sandwich (Target)		Pre-Cooked Meat (Target)		Dairy (NEDV)		Fresh Meat (NEDV)	
	χ^2	*p*-Value	χ^2	*p*-Value	χ^2	*p*-Value	χ^2	*p*-Value
Omnibus (Global)	187.41	<0.001 ***	131.76	<0.001 ***	17.87	0.764	92.93	<0.001 ***
Readability (Main)	0.14	0.713	0.33	0.567	0.33	0.566	0.36	0.551
Controls								
Sex	35.02	<0.001 ***	0.00	0.960	0.80	0.370	4.24	0.040 *
Age (Highest Sig.)	13.79	<0.001 ***	3.16	0.076	1.75	0.185	5.11	0.024 *
Religion	3.94	0.047 *	5.52	0.019 *	3.38	0.066	58.70	<0.001 ***
Income	13.18	<0.001 ***	17.43	<0.001 ***	2.01	0.156	1.10	0.294
Urban/Rural	5.43	0.020 *	0.59	0.442	0.01	0.943	0.40	0.526
Household Size	5.69	0.017 *	3.99	0.046 *	0.52	0.472	2.08	0.149
Marital Status	5.63	0.018 *	0.23	0.635	0.64	0.425	0.01	0.912
Year (2018)	9.71	0.002 **	9.10	0.003 **	1.29	0.257	8.02	0.005 **

Note: A significant p-value (< 0.05) suggests the parallel regression assumption is violated. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.2. Multicollinearity diagnostics (Variance Inflation Factors) for the predictors used in the study. Pre-packaged sandwich.

Variable	VIF	Df	Adj. VIF	VIF Equivalent
Readability (Q6_4)	1.19	1	1.09	1.19
Sex	1.22	1	1.10	1.22
Age	2.32	5	1.09	1.18
Religion	1.21	2	1.05	1.10
Income Tercile	1.26	2	1.06	1.12
Urban/Rural	1.08	1	1.04	1.08
Children (< 16)	2.72	1	1.65	2.72
Household Size	5.33	3	1.32	1.74
Marital Status	2.21	1	1.49	2.22
Ethnicity (Q3_1)	1.50	1	1.23	1.50
Country	1.17	2	1.04	1.08
Survey Year	1.09	3	1.02	1.03

Note: Diagnostics were calculated using a linear regression proxy, as standard VIF algorithms are not directly applicable to `polr` objects; however, multicollinearity is a property of the predictors, which remain identical across model specifications. All VIF equivalents are below the threshold of 5.

Table A.3. Binary logistic regression results: associations of perceived MTL readability and food consumption frequency (adjusted for all sociodemographic, behavioural, and temporal variables).

Variable	Pre-Packaged Sandwich		Pre-Cooked Meat		Dairy		Fresh Meat	
	β (SE)	**OR [95% CI]**	**β (SE)**	**OR [95% CI]**	**β (SE)**	**OR [95% CI]**	**β (SE)**	**OR [95% CI]**
Readability	−0.17 *** (0.05)	0.84 [0.77, 0.92]	0.09 (0.06)	1.09 [0.98, 1.22]	−0.09 (0.15)	0.92 [0.68, 1.23]	0.14 (0.09)	1.15 [0.96, 1.36]
Year (vs. 2012)								
2014	−0.20 (0.18)	0.82 [0.58, 1.16]	0.17 (0.25)	1.18 [0.73, 1.92]	0.25 (0.60)	1.28 [0.39, 4.17]	0.68 · (0.37)	1.97 [0.96, 4.04]
2016	0.11 (0.18)	1.11 [0.78, 1.58]	0.12 (0.25)	1.13 [0.70, 1.82]	−0.53 (0.65)	0.59 [0.17, 2.09]	−0.13 (0.34)	0.88 [0.45, 1.69]
2018	0.19 (0.18)	1.20 [0.84, 1.72]	−0.02 (0.25)	0.98 [0.61, 1.59]	−1.17 · (0.60)	0.31 [0.09, 1.01]	−0.45 (0.34)	0.64 [0.33, 1.23]
Interactions								
Readability × 2014	0.08 (0.06)	1.08 [0.96, 1.22]	−0.08 (0.09)	0.92 [0.77, 1.10]	−0.18 (0.20)	0.83 [0.56, 1.25]	−0.27 * (0.12)	0.76 [0.60, 0.97]
Readability × 2016	0.09 (0.06)	1.09 [0.96, 1.23]	−0.17 * (0.08)	0.84 [0.72, 0.99]	−0.05 (0.20)	0.95 [0.64, 1.41]	−0.17 (0.11)	0.84 [0.68, 1.05]
Readability × 2018	0.17 ** (0.06)	1.19 [1.05, 1.35]	−0.14 (0.08)	0.87 [0.74, 1.03]	0.03 (0.19)	1.03 [0.71, 1.49]	−0.01 (0.11)	0.99 [0.79, 1.23]

Note: Outcome is binary (Consumer = 1, Never = 0). Table displays coefficients (β) with standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). Significance levels: $\cdot\ p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.4. Reverse Causality: Ordinal logistic regression predicting perceived MTL readability from food consumption frequency (adjusted for all sociodemographic, behavioural, and temporal variables).

	Pre-Packaged Sandwich		**Pre-Cooked Meat**		**Dairy**		**Fresh Meat**	
Predictor	β (SE)	OR [95% CI]	β (SE)	OR [95% CI]	β (SE)	OR [95% CI]	β (SE)	OR [95% CI]
Consumption (vs. Never)								
Monthly	−0.27 * (0.12)	0.76 [0.60, 0.96]	0.34 * (0.17)	1.41 [1.00, 1.97]	−0.21 (0.53)	0.81 [0.29, 2.27]	0.23 (0.23)	1.26 [0.80, 1.97]
Weekly	−0.43 ** (0.14)	0.65 [0.49, 0.85]	0.24 (0.14)	1.27 [0.96, 1.68]	−0.04 (0.40)	0.96 [0.44, 2.10]	0.38 (0.21)	1.46 [0.97, 2.20]
Interactions								
Monthly × 2014	0.26 (0.15)	1.29 [0.96, 1.74]	−0.20 (0.25)	0.82 [0.51, 1.33]	−0.54 (0.69)	0.58 [0.15, 2.26]	−0.48 (0.30)	0.62 [0.34, 1.12]
Weekly × 2014	0.15 (0.20)	1.16 [0.78, 1.73]	−0.11 (0.19)	0.90 [0.61, 1.31]	−0.45 (0.51)	0.64 [0.23, 1.75]	−0.54 (0.29)	0.58 [0.33, 1.02]
Monthly × 2016	0.22 (0.16)	1.24 [0.91, 1.69]	−0.49 * (0.24)	0.61 [0.38, 0.98]	−0.27 (0.72)	0.76 [0.18, 3.12]	−0.32 (0.29)	0.73 [0.41, 1.28]
Weekly × 2016	0.21 (0.20)	1.23 [0.82, 1.83]	−0.54 ** (0.20)	0.58 [0.40, 0.86]	−0.53 (0.52)	0.59 [0.21, 1.63]	−0.49 (0.26)	0.61 [0.37, 1.02]
Monthly × 2018	0.34 * (0.16)	1.41 [1.03, 1.93]	−0.36 (0.23)	0.70 [0.44, 1.10]	0.11 (0.64)	1.12 [0.32, 3.96]	0.16 (0.28)	1.17 [0.67, 2.04]
Weekly × 2018	0.49 * (0.21)	1.64 [1.08, 2.47]	−0.30 (0.19)	0.74 [0.51, 1.08]	−0.20 (0.48)	0.82 [0.32, 2.12]	−0.11 (0.26)	0.90 [0.53, 1.51]

Note: Outcome variable is Perceived Readability. Table displays coefficients (β) with standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). Significance levels: * $p < 0.05$, ** $p < 0.01$.

Table A.5. Ordinal logistic regression results for ready meal category (2016).

	Ready Meals (2016)		
Variable	**β (SE)**	**OR**	**95% CI**
Readability	0.003 (0.040)	1.00	[0.93, 1.09]

Note: Table displays coefficients (β), standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). Model adjusted for all covariates in 2016.

Table A.6. Ordinal logistic regression results: associations of perceived MTL print size and REM consumption frequency (unadjusted).

	Pre-Packaged Sandwich			Pre-Cooked Meat		
Variable	**β (SE)**	**OR**	**95% CI**	**β (SE)**	**OR**	**95% CI**
Readability	−0.036 (0.023)	0.964	[0.921, 1.010]	−0.012 (0.023)	0.988	[0.944, 1.034]

Table displays coefficients (β), standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs).

ewpage

Table A.7. Mean and Standard Errors (SE) perceived readability of MTL labelling from 2012 to 2018, stratified by sociodemographic characteristics, behavioral characteristics, food products.

Variable	Category	2012	2014	2016	2018
Sex	Male	3.42 (0.05)	3.42 (0.05)	3.38 (0.05)	3.32 (0.05)
	Female	3.50 (0.06)	3.49 (0.05)	3.40 (0.06)	3.46 (0.06)
Age	16-24	3.91 (0.12)	4.03 (0.11)	3.90 (0.14)	4.23 (0.12)
	25-34	4.13 (0.07)	3.97 (0.08)	4.02 (0.08)	4.07 (0.07)
	35-44	2.87 (0.09)	2.95 (0.09)	2.87 (0.08)	2.91 (0.09)
	45-54	2.88 (0.07)	2.91 (0.07)	2.85 (0.06)	2.58 (0.05)
	55-64	3.93 (0.07)	3.82 (0.07)	3.90 (0.07)	3.93 (0.07)
	65+	3.11 (0.08)	3.21 (0.07)	3.16 (0.08)	3.18 (0.09)
Income	Low income	3.30 (0.06)	3.31 (0.05)	3.20 (0.07)	3.13 (0.06)
	High income	3.59 (0.05)	3.55 (0.05)	3.51 (0.05)	3.53 (0.05)
Religion	Christian	3.67 (0.07)	3.68 (0.05)	3.51 (0.06)	3.71 (0.06)
	Other religion	3.35 (0.05)	3.33 (0.05)	3.31 (0.05)	3.11 (0.05)
	No religion	3.60 (0.16)	3.33 (0.21)	3.54 (0.16)	3.67 (0.15)
Area	Urban	3.50 (0.04)	3.47 (0.04)	3.41 (0.05)	3.46 (0.05)
	Rural	3.21 (0.12)	3.32 (0.10)	3.32 (0.08)	3.17 (0.09)
Children at home	Yes	3.35 (0.05)	3.31 (0.05)	3.24 (0.05)	3.23 (0.05)
	No	3.75 (0.07)	3.78 (0.06)	3.75 (0.06)	3.75 (0.07)
Shopping resp.	All/Most	3.38 (0.05)	3.40 (0.04)	3.37 (0.05)	3.31 (0.05)
	Somet./Never	3.55 (0.06)	3.51 (0.05)	3.42 (0.06)	3.48 (0.06)
Country	England	3.46 (0.05)	3.46 (0.04)	3.39 (0.05)	3.39 (0.05)
	Wales	3.92 (0.09)	3.44 (0.12)	3.55 (0.09)	3.59 (0.09)
	N. Ireland	3.23 (0.16)	3.37 (0.07)	3.29 (0.11)	3.28 (0.08)
Concern safe eating	High	3.46 (0.05)	3.46 (0.04)	3.45 (0.04)	3.38 (0.05)
	Low	3.46 (0.08)	3.45 (0.09)	3.18 (0.10)	3.43 (0.08)
Information seeking	Yes	3.33 (0.09)	3.19 (0.09)	3.09 (0.09)	3.21 (0.09)
	No	3.50 (0.05)	3.50 (0.04)	3.46 (0.04)	3.44 (0.05)
Pre-packaged sandwich	Never	3.78 (0.09)	3.75 (0.09)	3.64 (0.09)	3.66 (0.09)
	Monthly	3.59 (0.06)	3.46 (0.06)	3.45 (0.06)	3.36 (0.06)
	Weekly	3.27 (0.06)	3.34 (0.05)	3.25 (0.05)	3.28 (0.07)

Table A.7. *Cont.*

Variable	Category	2012	2014	2016	2018
Pre-cooked meat	Never	3.45	3.46	3.44	3.38
		(0.05)	(0.05)	(0.05)	(0.05)
	Monthly	3.63	3.48	3.36	3.45
		(0.09)	(0.08)	(0.09)	(0.09)
	Weekly	3.36	3.42	3.33	3.37
		(0.08)	(0.08)	(0.07)	(0.07)
Diary	Never	3.47	3.46	3.40	3.39
		(0.04)	(0.04)	(0.04)	(0.04)
	Monthly	3.44	3.55	3.22	3.53
		(0.29)	(0.27)	(0.30)	(0.19)
	Weekly	3.30	3.00	3.20	3.36
		(0.29)	(0.23)	(0.25)	(0.19)
Fresh eat	Never	3.47	3.47	3.44	3.37
		(0.09)	(0.07)	(0.07)	(0.06)
	Monthly	3.44	3.47	3.38	3.35
		(0.05)	(0.05)	(0.05)	(0.05)
	Weekly	3.73	3.30	3.37	3.65
		(0.14)	(0.13)	(0.10)	(0.11)

Table A.8. Unweighted Attrition Analysis comparing the Analytic Sample ($N = 9,201$) versus Excluded Respondents ($N = 2688$).

Variable	**Excluded (N = 2688)**	**Included (N = 9201)**	***p*-Value**	**SMD**
Gender (%)			<0.001	0.106
Male	997 (37.1)	3891 (42.3)		
Female	1691 (62.9)	5310 (57.7)		
Age Group (%)			<0.001	0.325
16-24	315 (11.8)	582 (6.3)		
25-34	275 (10.3)	1404 (15.3)		
35-44	319 (11.9)	1588 (17.3)		
45-54	379 (14.2)	1660 (18.0)		
55-64	451 (16.9)	1513 (16.4)		
65+	933 (34.9)	2454 (26.7)		
Household Income Quartile (%)			0.031	0.352
Q1 (Low)	18 (24.7)	1440 (15.7)		
Q2	30 (41.1)	3098 (33.7)		
Q3	15 (20.5)	2657 (28.9)		
Q4 (High)	10 (13.7)	2006 (21.8)		
Urban/Rural Indicator (%)			0.029	0.049
Rural	521 (19.4)	1966 (21.4)		
Urban	2166 (80.6)	7235 (78.6)		
Children under 16 in HH (%)			<0.001	0.138
No	2128 (79.3)	6761 (73.5)		
Yes	554 (20.7)	2440 (26.5)		
Household Size (%)			<0.001	0.099
1 Person	769 (28.6)	2834 (30.8)		
2 People	911 (33.9)	3331 (36.2)		
3 People	422 (15.7)	1350 (14.7)		
4+ People	586 (21.8)	1686 (18.3)		
Country (%)			<0.001	0.113
England	1759 (65.4)	6479 (70.4)		
Wales	394 (14.7)	1241 (13.5)		
N. Ireland	535 (19.9)	1481 (16.1)		

Note: SMD = Standardized Mean Difference. SMD > 0.1 indicates imbalance.

Table A.9. Weighted Attrition Analysis. Comparison using Transportability Weights to adjust for sampling and non-response bias.

Variable	**Excluded** ***Weighted N (%)*** **(Total N = 2941)**	**Included** ***Weighted N (%)*** **(Total N = 8948)**	***p*-Value**	**SMD**
Gender			0.003	0.089
Male	1338.7 (45.5%)	4470.0 (50.0%)		
Female	1602.0 (54.5%)	4478.3 (50.0%)		
Age Group			<0.001	0.373
16-24	674.8 (23.1%)	995.4 (11.1%)		
25-34	430.8 (14.7%)	1556.5 (17.4%)		
35-44	343.3 (11.7%)	1541.9 (17.2%)		
45-54	399.3 (13.7%)	1763.9 (19.7%)		
55-64	387.2 (13.2%)	1232.8 (13.8%)		
65+	689.9 (23.6%)	1857.8 (20.8%)		
Household Income			0.228	0.308
Q1 (Low)	10.7 (19.2%)	920.7 (10.3%)		
Q2	17.8 (31.9%)	2549.9 (28.5%)		
Q3	12.5 (22.3%)	2833.2 (31.7%)		
Q4 (High)	14.9 (26.6%)	2644.5 (29.6%)		
Household Size			<0.001	0.240
1 Person	386.6 (13.1%)	1547.5 (17.3%)		
2 People	877.7 (29.8%)	3314.2 (37.0%)		
3 People	616.7 (21.0%)	1697.7 (19.0%)		
4+ People	1059.7 (36.0%)	2388.9 (26.7%)		
Urban/Rural			0.015	0.085
Rural	408.5 (13.9%)	1519.5 (17.0%)		
Urban	2531.0 (86.1%)	7428.8 (83.0%)		
Children (<16)			0.001	0.108
No	2204.6 (75.1%)	6291.9 (70.3%)		
Yes	730.9 (24.9%)	2656.4 (29.7%)		
Country			0.205	0.046
England	2668.0 (90.7%)	8232.4 (92.0%)		
Wales	171.6 (5.8%)	456.0 (5.1%)		
N. Ireland	101.1 (3.4%)	259.8 (2.9%)		

Note: Values are Weighted Counts (Column %). SMD < 0.1 suggests balance.

Table A.10. Ordinal logistic regression results: Associations of perceived MTL print size readability and food consumption frequency (Original 8-point scale) across product types. Adjusted for all sociodemographic and temporal variables.

	Sandwich (Target)		**Pre-Cooked Meat (Target)**		**Dairy (NEDV)**		**Fresh Meat (NEDV)**	
Variable	***β*** **(SE)**	**OR**	***β*** **(SE)**	**OR**	***β*** **(SE)**	**OR**	***β*** **(SE)**	**OR**
Readability	0.10 * (0.04)	1.11	0.00 (0.03)	1.00	0.07 (0.05)	1.07	−0.03 (0.04)	0.97
Year (ref: 2012)								
2014	0.32 (0.22)	1.37	0.02 (0.18)	1.02	0.01 (0.26)	1.01	−0.13 (0.21)	0.88
2016	0.41 (0.22)	1.51	−0.50 ** (0.17)	0.61	0.51 (0.27)	1.67	−0.71 *** (0.18)	0.49
2018	0.97 *** (0.21)	2.63	−0.38 * (0.18)	0.68	0.28 (0.25)	1.32	−0.80 *** (0.18)	0.45
Readability × Year								
Int: 2014	−0.07 (0.06)	0.93	−0.01 (0.05)	0.99	−0.04 (0.07)	0.96	0.05 (0.06)	1.05
Int: 2016	−0.06 (0.06)	0.94	0.06 (0.05)	1.06	−0.11 (0.07)	0.90	0.02 (0.05)	1.02
Int: 2018	−0.15 ** (0.06)	0.86	−0.02 (0.05)	0.98	−0.13 (0.07)	0.88	−0.01 (0.05)	0.99

Note: β = Unstandardized coefficient; SE = Standard Error; OR = Odds Ratio. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A.11. Weighted distribution of consumption frequency responses for the analytic sample ($N = 11,885$).

Product Type	Frequency	Weighted N	Percentage (%)
Target Products			
	Never/Rare	5417	45.6%
Pre-packaged Sandwiches	Monthly	4184	35.2%
	Weekly	2284	19.2%
	Never/Rare	2324	19.6%
Pre-cooked Meat	Monthly	2615	22.0%
	Weekly	6946	58.4%
Non-Equivalent Dependent Variables (NEDVs)			
	Never/Rare	352	3.0%
Dairy	Monthly	302	2.5%
	Weekly	11,231	94.5%
	Never/Rare	1187	10.0%
Fresh Meat	Monthly	2854	24.0%
	Weekly	7844	66.0%

Note: Weighted N rounded to nearest whole number. Percentages may not sum to 100 due to rounding.

Table A.12. Distribution of MTL print size and REM consumption levels across product types.

	Sandwich (Target)		**Pre-Cooked Meat (NEDV 1)**		**Fresh Meat (NEDV 2)**		**Dairy (NEDV 3)**	
Variable	β **(SE)**	**OR**	β **(SE)**	**OR**	β **(SE)**	**OR**	β **(SE)**	**OR**
Readability Level								
Level 2 (Difficult)	−0.47 * (0.24)	0.63	−0.03 (0.20)	0.97	−0.06 (0.24)	0.94	−0.12 (0.52)	0.89
Level 3 (Neither)	−0.29 (0.27)	0.75	0.03 (0.27)	1.03	−0.46 (0.35)	0.63	−0.96 (0.79)	0.38
Level 4 (Easy)	−0.69 ** (0.23)	0.50	−0.08 (0.22)	0.92	−0.06 (0.25)	0.94	−0.41 (0.54)	0.66
Level 5 (Very Easy)	−0.57 * (0.24)	0.57	0.14 (0.20)	1.15	0.16 (0.24)	1.17	−0.39 (0.55)	0.68
Test for Trend	Negative Gradient		No Trend (Null)		No Trend (Null)		No Trend (Null)	

oindentNote: β = Unstandardized coefficient; SE = Standard Error; OR = Odds Ratio (e^{β}). Reference category is Level 1 (Very Difficult). Significance levels: * $p < 0.05$, ** $p < 0.01$.

Table A.13. Comparison of the primary independent variable coefficient (Readability) between the Full Model (all controls) and a Parsimonious Model (excluding Religion, Household Size, Urban/Rural status, and Children under 16 at home).

Product	**Model Type**	**Readability β**	**% Change**	**Conclusion**
Pre-packaged Sandwich	Full	-0.098	0.3%	Robust (No Change)
	Parsimonious	-0.098		
Pre-cooked Meat	Full	0.030	48.7%	Robust (Null effect stable)
	Parsimonious	0.015		
Dairy	Full	-0.100	2.5%	Robust (No Change)
	Parsimonious	-0.097		
Fresh Meat	Full	0.054	3.7%	Robust (No Change)
	Parsimonious	0.052		

Note: Percent change is inflated due to the coefficient being near zero; the absolute difference ($\Delta = 0.015$) is negligible, and both models indicate a non-significant null result.

A.1.1 Ready-to Eat Meals Consumption Patterns

Table Table A.14 reveals updated temporal patterns in ready-to-eat meal consumption. For pre-packaged sandwiches, weekly consumption was relatively low but rose modestly over the six-year span, climbing from 17.3% in 2012 to 20.1% in 2018 (a 2.8-percentage-point increase). A larger shift occurred among occasional users: monthly consumption expanded from 33.4% to 42.4%, while the share of respondents who never ate pre-packaged sandwiches fell sharply from 49.3% to 37.5%. These changes indicate a movement away from complete avoidance toward at least occasional use of pre-packaged sandwiches. Pre-cooked meat products displayed the opposite trajectory. Weekly consumption started high at 66.4% in 2012 but declined steadily to 50.9% in 2018 (-15.5 percentage points). Meanwhile, monthly consumption nearly doubled—from 18.0% to 28.4% —and the proportion who never consumed pre-cooked meat inched up from 15.7% to 20.7%. Together, these figures suggest that some regular consumers of pre-cooked meats reduced the frequency of intake, shifting into the monthly or non-consumer categories by 2018.

To validate the Food and You patterns, we compared consumption trends with data from the National Diet and Nutrition Survey (NDNS). While NDNS does not include identical product categories, it captures consumption of manufactured meat products and ready meals (including beef, chicken, pork, and fish-based products) measured. To estimate the annual prevalence of ready-to-eat meal consumption in the NDNS, we first identified all relevant ready-to-eat or manufactured products (see Appendix A.1.1.1). For each survey participant, we summed the number of ready-made products reported consumed during the 4-day period. Individuals reporting at least one ready meal were classified as consumers. To ensure national representativeness, we calculated the weighted annual prevalence of ready meal consumption for each survey year.

The comparison reveals several key convergences that support external validity. First, temporal trends align consistently across datasets: NDNS ready meal consumption increased from 66.7% to 70.0% (2012-2018), paralleling the FYS's documented shift toward

Table A.14. Comparative analysis of REM consumption trends: Food and You and NDNS Data (No survey weights)

Food and You Survey	Category $n = 8954$	2012 $n = 2136$	2014 $n = 2336$	2016 $n = 2309$	2018 $n = 2173$
Pre-packaged sandwich	Weekly	369 (17.3%)	455 (19.5%)	415 (18.0%)	437 (20.1%)
	Monthly	713 (33.4%)	741 (31.7%)	849 (36.8%)	921 (42.4%)
	Never	1054 (49.3%)	1141 (48.8%)	1045 (45.3%)	815 (37.5%)
Pre-cooked meat	Weekly	1418 (66.4%)	1505 (64.4%)	1271 (55.0%)	1107 (50.9%)
	Monthly	384 (18.0%)	446 (19.1%)	591 (25.6%)	616 (28.4%)
	Never	334 (15.7%)	386 (16.5%)	447 (19.4%)	450 (20.7%)
NDNS	$n = 3019$	$n = 795$	$n = 764$	$n = 758$	$n = 702$
Ready-to-eat meals	Annually	531 (66.7%)	468 (61.2%)	505 (67.0%)	491 (70.0%)

more frequent convenience food consumption. Specifically, weekly and monthly pre-cooked meat consumption rose indicating growing acceptance of manufactured meat products that mirrors NDNS trends. Second, the magnitude of consumption patterns demonstrates comparable population engagement with convenience foods. NDNS data showing approximately 67% annual ready meal consumption aligns with Food and You findings that 33.4% of respondents consume pre-packaged sandwiches at least monthly in 2012, increasing to 42.4%. Third, both surveys capture similar demographic representativeness with substantial sample sizes (FYS: n=8854; NDNS: n=3019) and consistent year-over-year participation, lending confidence to trend analyses.

A.1.1.1 Ready-to-Eat Meals List in NDNS 2008-2019

- MANUFACTURED BEEF PRODUCTS INCLUDING READY MEALS
- MANUFACTURED CANNED TUNA PRODUCTS INCL READY MEALS

- MANUFACTURED CHICKEN PRODUCTS INCL READY MEALS
- MANUFACTURED COATED CHICKEN / TURKEY PRODUCTS
- MANUFACTURED EGG PRODUCTS INCLUDING READY MEALS
- MANUFACTURED LAMB PRODUCTS INCLUDING READY MEALS
- MANUFACTURED MEAT PIES AND PASTRIES
- MANUFACTURED OILY FISH PRODUCTS INCL READY MEALS
- MANUFACTURED PORK PRODUCTS INCLUDING READY MEALS
- MANUFACTURED SHELLFISH PRODUCTS INCL READY MEALS
- MANUFACTURED WHITE FISH PRODUCTS INCL READY MEALS
- MEAT ALTERNATIVES INCL READY MEALS & HOMEMADE DISH
- OTHER MEAT PRODUCTS MANUFACTURED INCL READY MEALS
- READY MEALS BASED ON SAUSAGES

A.1.1.2 Transportability Analysis

To validate and evaluate the applicability of our primary findings, we conduct a non-nested transportability analysis [32]. This analysis adjusts the results from a study sample to align with a target population by accounting for differences in baseline characteristic distributions [92]. Although similar weighting techniques have recently been employed to transport causal estimates from randomized trials to representative survey populations, we modified these techniques to assess the consistency of a statistical association between two distinct national representative surveys [183, 184]. In this study, the application of transportability methods is justified by the inherent differences in diet measurement and sampling methodologies employed in nutritional surveys. Food and You survey served

as the study sample, characterized by a single visit including questions on attitudes and perceptions, whereas the target population was drawn from the NDNS, which utilizes a exhaustive 4-day food diary [56, 156]. To meet the positivity assumption necessary for transportability, the NDNS sample excluded participants from Scotland, and included only adults who consumed at least one ready-to-eat meal product between 2012 and 2019 (see Appendixes Section A.1 and Section A.1.1). After applying these exclusion criteria, the final NDNS sample included 1988 participants.

Table A.15. Covariate distributions between the Food & You and NDNS Samples. All covariates included in the participation model

Characteristic	**Level**	**Food & You (Unweighted) (N = 8949)**	**NDNS Target (Weighted) (N = 1988)**
Sex			
	Male	4470 (49.9%)	853 (42.9%)
	Female	4479 (50.1%)	1135 (57.1%)
Age Group			
	16–24	996 (11.1%)	201 (10.1%)
	25–34	1556 (17.4%)	302 (15.2%)
	35–44	1542 (17.2%)	335 (16.9%)
	45–54	1764 (19.7%)	352 (17.7%)
	55–64	1232 (13.8%)	311 (15.6%)
	65+	1859 (20.8%)	487 (24.5%)
Household Size			
	1	1548 (17.3%)	512 (25.8%)
	2	3315 (37.0%)	697 (35.1%)
	3	1697 (18.9%)	304 (15.3%)
	4+	2389 (26.7%)	475 (23.9%)
Marital Status			
	Married/Similar	5294 (59.2%)	923 (46.4%)
	Single/Similar	3655 (40.8%)	1065 (53.6%)

To interpret the results derived from transportability analyses as indicators of the MTL readability's effectiveness within the NDNS population, we establish two primary assumptions. Firstly, we presume the transportability of MTL readability's effects, conditional on the observed covariates, across differing Food and You participation statuses [32, 184]. Essentially, this assumption posits that the covariates incorporated into the model for predicting Food and You participation are adequate for adjusting for selective participation

Table A.15. *Cont.*

Characteristic	Level	Food & You (Unweighted) (N = 8949)	NDNS Target (Weighted) (N = 1988)
Children at home			
	No	6291 (70.3%)	1623 (81.6%)
	Yes	2658 (29.7%)	365 (18.4%)
Income Tercile			
	Lowest	4539 (49.3%)	509 (25.6%)
	Middle	2657 (28.9%)	557 (28.0%)
	Highest	2006 (21.8%)	682 (34.3%)
	Missing	0 (0.0%)	240 (12.1%)
Country			
	England	8232 (91.9%)	1461 (73.5%)
	Wales	456 (5.1%)	286 (14.4%)
	N. Ireland	261 (2.9%)	241 (12.1%)

Note: The NDNS target sample is filtered to include adults from England, Wales, and Northern Ireland who consumed REM between 2012–2019.

in the survey [43]. Secondly, we assume that every combination of these covariates present in the NDNS population has a non-zero probability of being represented within the Food and You sample [32].

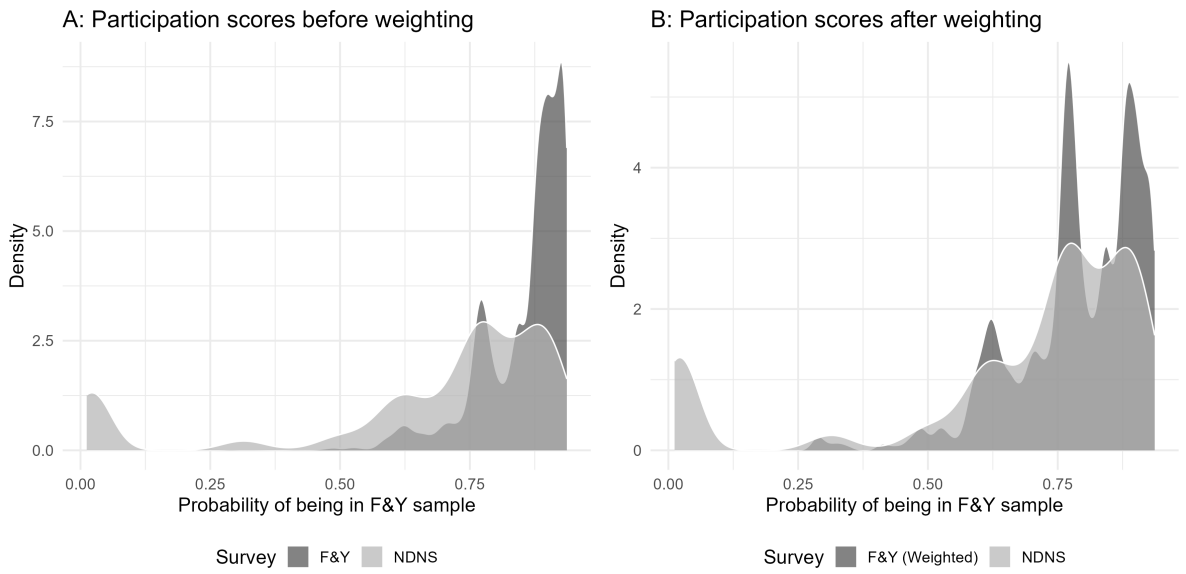

Fig. A.1. Distribution of the estimated participation scores for the Food & You (F&Y) and National Diet and Nutrition Survey (NDNS) samples. Panel (**A**) shows the distributions before weighting, and Panel (**B**) shows the distributions after applying the inverse probability of sampling weights to the F&Y sample. The improved overlap in Panel (**B**) supports the positivity assumption for the transportability analysis.

With a comparable target population defined, a three-step weighting methodology was used for the transportability analysis. Initially, the partial proportional odds ordinal logistic regression models to estimate the primary association within the study sample, accounting for the Food and You survey's complex design [43]. Following this, a participation model was constructed using a Generalized Boosted Model (GBM), a machine learning technique adept at handling intricate, non-linear relationships without predefined assumptions [141, 43]. This model was trained on combined datasets to predict survey membership based on a suite of harmonized, shared covariates. The following baseline covariates were included in the participation model for adjustment: sex, age, children under 16 in the household, household size, marital status, household income and country. Other variables, such as shopping responsibilities or health beliefs, were not available in the NDNS database; however, these characteristics were measured in the Food and You database and thus were not included for the participation model. The final step involved calculating Inverse Probability of Sampling Weights (IPSW) using the predicted probabilities from the GBM,

Table A.16. Sensitivity Analysis: Comparison of transported associations using different covariate sets in the participation model

	Transported (Demographics Model)		Transported (Fully Adjusted Mode	
Variable	**β (SE)**	**OR [95% CI]**	**β (SE)**	**OR [95% CI]**
Pre-packaged Sandwich				
Readability	−0.11 * (0.05)	0.89 [0.81, 0.99]	−0.21 ** (0.07)	0.81 [0.70, 0.92]
Year (vs. 2012)				
2014	−0.38 (0.28)	0.68 [0.39, 1.18]	−0.59 * (0.31)	0.55 [0.30, 1.01]
2016	−0.40 (0.26)	0.67 [0.40, 1.12]	−0.92 ** (0.32)	0.40 [0.21, 0.74]
2018	−0.94 *** (0.26)	0.39 [0.23, 0.66]	−1.36 *** (0.31)	0.26 [0.14, 0.47]
Interactions				
Readability × 2014	0.07 (0.08)	1.07 [0.92, 1.25]	0.13 (0.09)	1.14 [0.95, 1.36]
Readability × 2016	0.05 (0.07)	1.05 [0.91, 1.21]	0.17 * (0.09)	1.19 [1.00, 1.42]
Readability × 2018	0.13 (0.07)	1.13 [0.98, 1.30]	0.23 ** (0.09)	1.26 [1.06, 1.50]
Pre-cooked Meat				
Readability	0.08 (0.05)	1.08 [0.98, 1.19]	0.01 (0.06)	1.01 [0.90, 1.14]
Year (vs. 2012)				
2014	0.28 (0.26)	1.33 [0.80, 2.20]	−0.23 (0.32)	0.80 [0.42, 1.49]
2016	0.85 *** (0.24)	2.35 [1.47, 3.75]	0.67 ** (0.29)	1.96 [1.11, 3.46]
2018	0.87 *** (0.23)	2.39 [1.53, 3.74]	0.49 * (0.28)	1.62 [0.94, 2.80]
Interactions				
Readability × 2014	−0.08 (0.07)	0.93 [0.81, 1.07]	0.05 (0.08)	1.05 [0.89, 1.23]
Readability × 2016	−0.12 * (0.07)	0.89 [0.78, 1.01]	−0.09 (0.08)	0.91 [0.78, 1.07]
Readability × 2018	−0.08 (0.07)	0.92 [0.81, 1.05]	0.03 (0.08)	1.03 [0.88, 1.19]

Note: Table displays coefficients (β) with standard errors (SEs), odds ratios (ORs), and 95% confidence intervals (CIs). The "Demographics Model" adjusted for age and sex. The "Fully Adjusted Model" additionally adjusted for household size, children, marital status, country, and income tercile. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

which were then multiplied by the original Food and You survey weights to generate a comprehensive, combined weight [92, 185]. The initial ordinal logistic regression models were subsequently refitted with this combined weight to yield the transported estimates.

Chapter B

Appendix Chapter 4

Table B.1. Calorie count for breakfast cereals included in the choice experiment ($n = 8$)

Cereal brand	Calorie count (kcal per 100 g)	Coarse	Detailed
None option	0	-	-
1	360	Very low	A
2	374	Very low	B
3	392	Low	C
4	398	Low	D
5	423	High	E
6	431	High	F
7	453	Very high	G
8	470	Very high	H

Table B.2. Participant characteristics by experimental group ($n = 498$)

Variable	Alternative	All	Absent n (%)	Coarse n (%)	Detailed n (%)	χ^2 (p)
Participants		498	162 (32)	167 (34)	169 (34)	
Sex	Female	257 (52)	86 (54)	91 (55)	80 (47)	0.39
Sex	Male	240 (48)	75 (46)	75 (45)	89 (53)	
Sex	Unreported	1 (0)	1 (0)			
Age	18-34	140 (28)	44 (27)	46 (28)	50 (29)	0.94
Age	35-54	165 (33)	55 (34)	58 (35)	52 (31)	
Age	55-65+	193 (39)	63 (39)	63 (37)	67 (40)	
Ethnicity	White	417 (84)	138 (85)	139 (83)	140 (83)	0.89
Ethnicity	Mixed	75 (15)	23 (14)	26 (16)	26 (15)	
Ethnicity	Unreported	6 (1)	1 (1)	2 (1)	3 (2)	
Education	Higher	199 (40)	59 (36)	66 (40)	74 (44)	0.57
Education	Lower	297 (60)	102 (63)	101 (60)	94 (56)	
Education	Unreported	2 (0)	1 (1)	0 (0)	1 (1)	
Income[1]	Higher	267 (54)	93 (57)	96 (58)	78 (46)	0.13
Income	Lower	195 (39)	60 (37)	57 (34)	78 (46)	
Income	Unreported	36 (7)	9 (6)	14 (8)	13 (8)	
BMI [2]	Obesity	130 (26)	45 (28)	38 (23)	47(28)	0.95
BMI	Overweight	159 (32)	52 (32)	55 (33)	52 (31)	
BMI	Normal weight	191 (38)	59 (36)	68 (41)	64(38)	
BMI	Underweight	18 (4)	6 (4)	6 (4)	6 (4)	

[1] Higher income: £30,001 - Above £40,000; Lower income: Below £10,000 - £30,000. [2] BMI (body mass index) was estimated with self-reported height and weight.

Table B.3. Household composition and purchasing behaviors by experimental group ($n = 498$)

Variable	Alternative	All	Absent n (%)	Coarse n (%)	Detailed n (%)	χ^2 (p)
Participates		498	162 (32)	167 (34)	169 (34)	
Children at home	No	334 (67)	103 (64)	108 (65)	123 (73)	0.39
Children at home	Yes	157 (32)	56 (35)	57 (34)	44 (26)	
Children at home	Unreported	7 (1)	3 (1)	2 (1)	2 (1)	
Losing weight	No	224 (45)	70 (43)	75 (45)	79 (47)	0.47
Losing weight	Yes	268 (54)	91 (56)	91 (54)	86 (51)	
Losing weight	Unreported	6 (1)	1 (1)	1 (1)	4 (2)	
Responsibility	Always	255 (51)	83 (51)	87 (52)	85 (50)	0.99
Responsibility	Most of the time	135 (27)	45 (28)	45 (27)	45 (27)	
Responsibility	Half of the time	60 (12)	18 (11)	20 (12)	22 (13)	
Responsibility	Sometimes	43 (9)	15 (09)	13 (07)	15 (08)	
Responsibility	Never	5 (1)	1 (1)	2 (1)	2 (1)	
Check ingredients	Always	58 (12)	23 (14)	20 (12)	15 (09)	0.07
Check ingredients	Most of the time	177 (36)	58 (36)	70 (42)	49 (29)	
Check ingredients	Half of the time	111 (22)	31 (19)	36 (22)	44 (26)	
Check ingredients	Sometimes	132 (27)	43 (27)	33 (20)	56 (33)	
Check ingredients	Never	20 (4)	7 (4)	8 (8)	5 (3)	
Familiarity	Very familiar	174 (35)	52 (32)	70 (41)	52 (31)	0.20
Familiarity	Somewhat familiar	272 (55)	90 (56)	82 (50)	100 (60)	
Familiarity	Neither	20 (4)	6 (04)	5 (03)	9 (05)	
Familiarity	Unfamiliar	32 (6)	14 (8)	10 (6)	8 (4)	

Table B.4. Product preferences by experimental group, means and SDs ($n = 498$)

Cereal brand	Calorie count	Absent n (%)	Coarse n (%)	Detailed n (%)	Anova (p)
Participates		162 (0.32)	167 (0.34)	169 (0.34)	
1	360	4.31 (2.37)	4.07 (2.42)	3.98 (2.36)	0.43
2	374	3.49 (2.49)	3.39 (2.56)	3.91 (2.74)	0.15
3	392	4.98 (1.83)	4.55 (2.04)	4.66 (1.93)	0.12
4	398	4.31 (1.94)	3.75 (1.95)	3.93 (2.00)	0.03
5	423	5.47 (2.03)	5.32 (1.97)	5.31 (2.17)	0.74
6	431	3.61 (1.88)	4.26 (1.86)	3.96 (1.79)	0.00
7	453	5.94 (2.37)	6.04 (2.24)	5.76 (2.37)	0.54
8	470	3.88 (2.13)	4.60 (2.07)	4.49 (2.11)	0.00

Table B.5. *d'* scores by experimental group, n (Mean, SD)

Variable	1-back (n=498)	2-back (n=464)	3-back (n=452)
Absent	162 (3.37, 1.01)	157 (3.24, 1.00)	150 (2.29, 0.87)
Coarse	167 (3.13, 1.26)	150 (3.24, 0.91)	148 (2.29, 0.82)
Detailed	169 (3.21, 1.17)	157 (3.14, 0.93)	154 (2.16, 0.90)
Anova (p)	0.16	0.59	0.30

Table B.6. Subjective FOP understanding questions by experimental group (n = 498)

Variable	Alternative	All	Absent	Coarse	Detailed	χ^2 (p)
Participants		498	162 (32)	167 (34)	169 (34)	
Reported seeing FOP	Always	147 (44)		80 (48)	67 (40)	0.22
Reported seeing FOP	Most of the time	104 (31)		51 (36)	53 (31)	
Reported seeing FOP	About half the time	32 (10)		10 (6)	22 (13)	
Reported seeing FOP	Sometimes	39 (11)		19 (11)	20 (12)	
Reported seeing FOP	Never	14 (4)		7 (4)	7 (4)	
Having information	Overload or sufficient	244 (49)	54 (33)	99 (60)	91 (54)	0.00
Neither		66 (13)	26 (16)	19 (11)	21 (12)	
Somewhat insufficient		133 (27)	51 (31)	36 (22)	46 (27)	
Insufficient		55 (11)	31 (19)	13 (8)	11 (7)	

Chapter C

Appendix Chapter 5

Nutritional label randomized control trial: Randomization and consort diagram

Subjects were randomized to one of three experimental conditions: Control (n=162), Coarse (n=167), or Detailed (n=169). The labels, adapted from prior research, indicated caloric content per 100g for a range of breakfast cereals (350–478 kcal). The Coarse condition employed four categories: very low (350–382 kcal), low (383–414 kcal), high (415–446 kcal), and very high (447–478 kcal). The Detailed condition employed an eight-category scale, with each letter corresponding to a specific calorie range: A (350–366 kcal), B (367–382 kcal), C (383–398 kcal) through H (463–478 kcal).

In each trial of the 16-trial choice experiment, participants were presented with a set of four cereal brands and asked to select the one they would most likely purchase. The nutritional data for these cereals ranging from 360 to 470 kcal were sourced from manufacturer websites and verified against UK supermarket databases. To mitigate positional and order bias, the spatial arrangement of cereals within each trial and the sequence of the 16 trials were randomized for every participant.

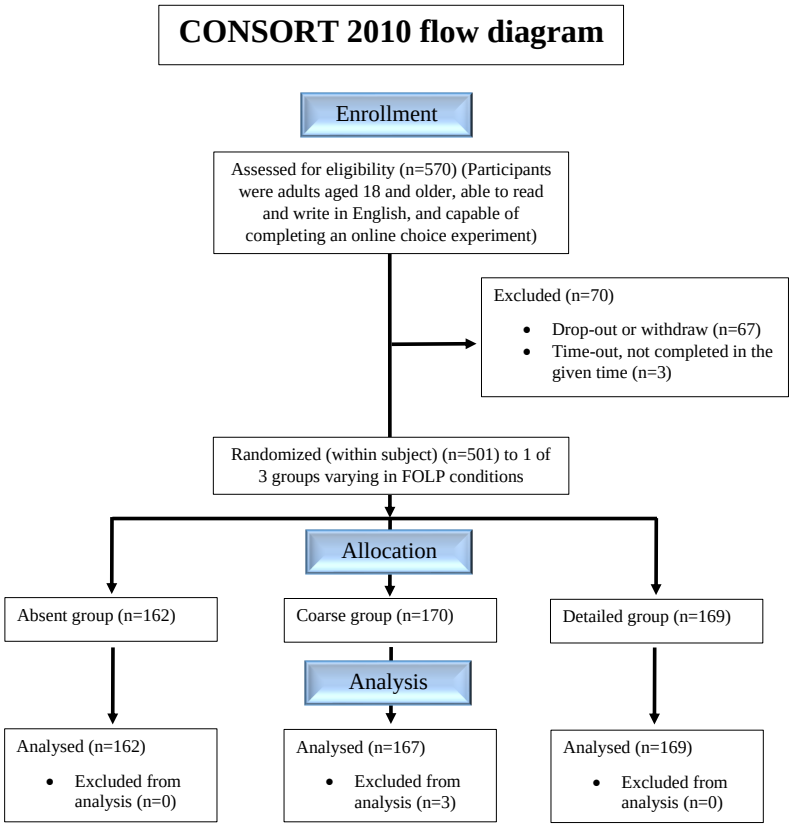

Fig. C.1. CONSORT flow diagram illustrating the recruitment, randomization, and follow-up of participants in the food label granularity trial.

Transportability assumption checks

Fig. C.2. Diagnostic plot for positivity. The mirrored density plot displays the distribution of the estimated propensity scores—defined as the probability of trial participation $P(S = 1|X)$—for the source population (RCT, top panel) and the target population (NDNS, bottom panel). The substantial overlap indicates that no sub-populations in the target group are systematically excluded, supporting the validity of the transportability weights.

Collinearity analysis: Variance Inflation Factors (VIF)

Table C.1. Collinearity analysis: Variance Inflation Factors (VIF) for RCT Data

Variable	GVIF	Df	Adjusted GVIF
Age	1.613	5	1.049
Sex	1.068	1	1.033
Ethnicity	1.247	4	1.028
Country	1.114	3	1.018
Trying to lose weight	1.168	1	1.081
BMI	1.174	1	1.084
Children in household	1.210	1	1.100
Education level	1.291	3	1.044
d-prime (2-back)	1.579	1	1.257
d-prime (3-back)	1.649	1	1.284

Note: Adjusted GVIF represents $GVIF^{1/(2 \cdot Df)}$. Values close to 1 indicate a lack of multicollinearity.

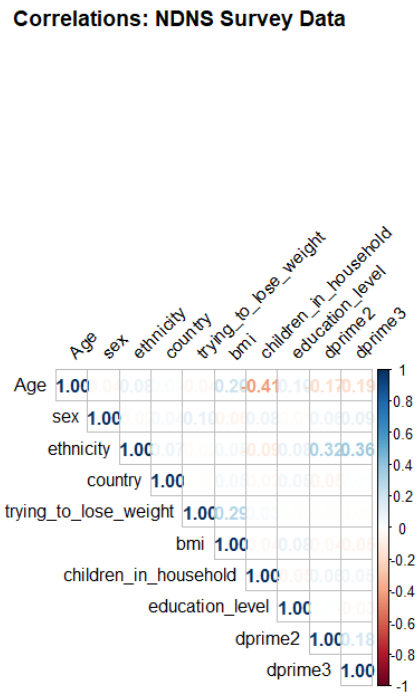

Fig. C.3. Pairwise correlation matrix for sociodemographic, behavioral, and cognitive covariates in the RCT dataset. Blue indicates positive and red indicates negative correlations. No pairwise correlation coefficient exceeded 0.8, providing evidence alongside VIFs that multicollinearity does not significantly threaten the stability of the regression estimates.

Table C.2. Collinearity analysis: Variance Inflation Factors (VIF) for NDNS Survey Data

Variable	GVIF	Df	Adjusted GVIF
Age	1.781	5	1.059
Sex	1.043	1	1.021
Ethnicity	1.398	2	1.087
Country	1.126	3	1.020
Trying to lose weight	1.120	1	1.058
BMI	1.165	1	1.079
Children in household	1.401	1	1.184
Education level	1.181	3	1.028
d-prime (2-back)	1.203	1	1.097
d-prime (3-back)	1.342	1	1.158

Note: Values are calculated on the analytical NDNS sample ($N = 10,696$).

Estimated PATE using GLM-IPSW

Table C.3. Sensitivity Analysis: Estimated PATE using GLM-IPSW

Covariate Specification	Coarse vs Control	Detailed vs Control	Coarse vs Detailed
Model A (Demographics)	0.3 [-24.9, 23.0]	1.3 [-23.1, 17.7]	-1.0 [-20.0, 10.2]
Model B (+ Lifestyle)	12.8 [-19.7, 39.9]	13.6 [-14.3, 36.4]	-0.8 [-20.1, 18.1]
Model C (+ SES)	1.2 [-26.6, 24.6]	3.2 [-20.2, 21.8]	-2.0 [-20.3, 18.9]
Model D (+ Working Memory $d'2$)	11.9 [-18.2, 43.2]	19.6 [3.3, 41.7]*	-7.7 [-28.8, 9.7]
Model E (+ Working Memory $d'3$)	-0.7 [-22.5, 21.8]	11.1 [-5.5, 30.3]	-11.8 [-28.7, 2.5]

Note: Estimates represent the change in daily calorie intake (kcal). Values in brackets are 95% confidence intervals derived from 50 survey-weighted bootstrap iterations. In this sensitivity analysis, the probability of trial participation was estimated using a parametric logistic regression model (GLM-IPSW) incorporating complex survey weights, rather than the BART used in the main analysis. Asterisks (*) denote statistical significance (95% CI excludes zero). Model A: Age, sex, ethnicity, country. Model B: Model A + BMI, intent to lose weight, children in household. Model C: Model B + Education Level. Model D: Model C + Working Memory ($d'2$). Model E: Model C + Working Memory ($d'3$).

Estimated PATE using TMLE-BART across linear and bound synthetic specifications

Table C.4. Comparison of nutritional label effects on caloric count: sensitivity analysis from OLS models versus transported PATE via DR TMLE-BART

Model	N_{RCT}	N_{NDNS}	Coarse vs Control	Detailed vs Control	Coarse vs Detailed
Naive (OLS Unadjusted)					
Naive (Mean Diff)	490	–	-9.1 [-23.4, 5.1]	-3.5 [-18.1, 11.2]	-5.7 [-19.0, 7.6]
Naive (OLS Adjusted - Interaction Effects)					
Naive Interaction ($d'2$)	464	–	-15.2 [-30.1, -0.3]	-19.2 [-33.8, -4.6]	4.0 [-16.8, 24.8]
Naive Interaction ($d'3$)	452	–	-10.2 [-27.3, 6.9]	-17.3 [-33.4, -1.2]	7.1 [-16.4, 30.6]
Transported PATE (DR TMLE-BART) - Sensitivity Bounds: Model E (Adjusted for $d'2$)					
Model E (Lower -10%)	464	10,696	-3.7 [-32.7, 26.9]	27.6 [4.8, 53.8]	-31.2 [-52.3, -9.8]
Model E (Upper +10%)	464	10,696	-9.2 [-36.5, 18.1]	15.2 [-8.6, 39.0]	-24.4 [-46.0, -3.4]
Model E (Linear)	464	10,696	-5.6 [-32.3, 23.1]	22.5 [0.4, 48.3]	-28.1 [-47.7, -7.9]
Transported PATE (DR TMLE-BART) - Sensitivity Bounds: Model F (Adjusted for $d'3$)					
Model F (Lower -10%)	452	10,696	-4.1 [-31.9, 23.3]	20.8 [-0.5, 43.1]	-24.9 [-43.9, -8.0]
Model F (Upper +10%)	452	10,696	-7.2 [-31.2, 15.1]	14.6 [-7.3, 36.2]	-21.8 [-41.3, -4.1]
Model F (Linear)	452	10,696	-7.7 [-31.5, 16.0]	16.2 [-2.9, 37.8]	-23.9 [-43.4, -6.2]

Note: Naive estimates: comparisons among treatment groups derived from RCT data. Naive Interaction: estimates evaluated at working memory capacity means (d'). DR TMLE-BART: transported PATE on calorie count for the NDNS population, with 95% CIs from 500 bootstrap iterations. Generative models E and F incorporate alternative structural assumptions including deterministic ±10% shifts for cognitive upper/lower bounds.

Estimated PATE using TMLE-BART by years

Year	Model D (High $d'2$ Adjusted) *Est [95% CI]*	Model E (High $d'3$ Adjusted) *Est [95% CI]*
2008	-21.6 [-42.1, -3.0]*	-27.9 [-49.4, -12.7]*
2009	-20.2 [-36.1, -7.2]*	-21.6 [-36.7, -3.9]*
2010	-23.2 [-37.5, -5.9]*	-21.2 [-39.6, -5.9]*
2011	-21.1 [-39.3, 0.4]	-24.8 [-40.0, -7.4]*
2012	-21.8 [-37.9, -4.2]*	-21.5 [-38.6, -2.7]*
2013	-19.8 [-34.1, -1.9]*	-25.3 [-40.1, 2.1]
2014	-14.7 [-31.7, 3.6]	-15.9 [-39.4, 7.1]
2015	-8.8 [-30.1, 11.2]	-17.7 [-36.4, 1.9]
2016	-10.7 [-27.9, 8.6]	-15.3 [-37.2, 6.9]
2017	-11.4 [-32.1, 3.4]	-17.0 [-37.1, 0.4]
2018	-13.4 [-36.1, 5.9]	-13.8 [-34.4, 5.9]
Pandemic Period		
2019	-19.2 [-46.1, -1.8]*	-23.1 [-34.2, -10.5]*
2020	-16.8 [-31.3, 2.9]	-19.3 [-33.5, -5.5]*
2021	-17.0 [-37.5, -1.4]*	-19.4 [-35.6, -1.7]*
2022	-14.4 [-33.3, -1.5]*	-19.2 [-40.5, -8.7]*

Note: Time-Series Sensitivity Analysis (2008–2022). Estimated PATE for the Coarse vs. Detailed comparison across individual survey years derived from separate DR-BART models.

Synthetic data generation for d' values: technical note

Overview and causal justification

The transportability analysis of the PATE required the inclusion of working memory, measured via standard signal detection scores, as a critical pre-treatment effect modifier. However, these cognitive metrics were absent from the national target survey. To address this missing covariate problem while maintaining structural integrity, we developed a five-step generative machine learning and DR bootstrapping framework.

Step 1: Global multiple imputation

To address baseline missingness in the cross-sectional target survey prior to synthesis, we utilized MICE using a random forest engine. This approach captures non-linearities and interactions in epidemiological data. The imputation model included demographics,

lifestyle variables, BMI, household composition, education, and survey year. The inclusion of the survey year accounted for temporal variations and secular trends across repeated cross-sectional data spanning 2008 to 2023. Missingness rates for behavioral variables and BMI were 6.2 percent and 11.4 percent, respectively. We generated five imputed datasets, pooling final causal point estimates and standard errors using standard combination rules. Cognitive variables were detached during this step to prevent data leakage.

Table C.5. Missingness statistics for NDNS covariates (N=10,696)

Variable	N Missing	% Missing
Age	0	0.0%
Sex	0	0.0%
Ethnicity	82	0.8%
BMI	1,219	11.4%
Trying to lose weight	663	6.2%
Children in household	150	1.4%
Education level	342	3.2%
Working memory (d-prime)	10,696	100.0%*

**Note: Working memory tests were not administered in the NDNS survey and were fully imputed based on demographic profiles.*

Table C.6. Missing Data Diagnostics: Predictors of Missingness in NDNS Survey Data

Significant Predictor	Test Method	P-Value
Country	Chi-Square	< 0.001
Working memory (3-back)	T-Test	< 0.001
Working memory (2-back)	T-Test	< 0.001
Age	Chi-Square	< 0.001

Note: This table displays significant associations between the presence of missing data and observed covariates, supporting the Missing At Random (MAR) assumption for MICE.

Step 2: Synthetic data generation

Given the total absence of cognitive scores in the target population, we utilized a synthetic data generation framework. We implemented a dual engine approach trained on the experimental trial population. The primary engine utilized a gradient boosting framework, while the secondary engine utilized a piecewise linear regression to prevent algorithmic

dependency. The models learned the conditional distribution of cognitive scores given covariates including age, sex, ethnicity, education level, BMI, and survey year. Base predictions were refined using a penalty function derived from longitudinal aging trends. The penalty was applied based on chronological age (A):

$$\Delta_{penalty} = \begin{cases} 0, & \text{if } A < 55 \\ (A - 54) \times -0.01125, & \text{if } 55 \leq A < 75 \\ -0.225 + ((A - 74) \times -0.054), & \text{if } A \geq 75 \end{cases}$$

An additional penalty of -0.15 was applied for chronological age 75 or older with low education. Biological heterogeneity was reintroduced via a stochastic shock ($\epsilon \sim \mathcal{N}(0, 0.5)$).

Step 3: Weighted bootstrapping

To represent the national population across the temporal span, we implemented a weighted bootstrapping procedure. Survey weights corrected for unequal selection probabilities and non-response. Temporal fractions adjusted the influence of each survey year. The target dataset was resampled with replacement using adjusted global weights to propagate sampling variance throughout the pipeline.

Step 4: DR estimation and sensitivity analysis

We estimated the target parameter using TMLE powered by BART. This semiparametric, DR framework optimizes the bias-variance tradeoff by initially estimating outcome and participation models, followed by a targeted fluctuation step. This approach guarantees consistency if either model is correctly specified.

To assess structural uncertainty, we nested sequential model specifications within every iteration, adjusting for demographics (Model A), household (Model B), lifestyle (Model

C), and socioeconomic status (Model D). Finally, Model E and Model F introduced synthetic working memory scores. We conducted a parameter bound sensitivity analysis by evaluating results across deterministic bounds shifted by a 10 percent margin in both directions.

Step 5: High performance computing protocol

The pipeline was deployed on university supercomputer facilities to handle computational demands. The 500-iteration bootstrap was parallelized into independent tasks with jobs routed to high-memory partitions (32GB RAM per node). Single threading was enforced to prevent thread contention crashes in machine learning loops, ensuring stable execution of the multi-faceted sensitivity analysis.

Validation and uncertainty propagation

Distributional similarity was validated by comparing moments of the trial and target populations. The correlation between age and cognitive decline confirm that applied penalties maintained biological trends. The entire pipeline was embedded within the bootstrap loop to incorporate both sampling and generative structural variance in the final results.

Dynamic microsimulation: technical note

Methodological framework and cohort construction

The microsimulation utilizes a dynamic, stochastic, individual-level architecture, classified as a discrete-time state-transition model. The baseline virtual cohort was constructed using NDNS records spanning 2008 to 2023. We implemented temporal pooling to preserve

national representativeness, scaling the cohort to represent 53 million adults in the UK. Baseline data for BMI, education, and ethnicity were imputed using gradient-boosted trees.

Physiological dynamics and energy balance

Body weight transitions were modeled using dynamic energy balance equations. We adopted a framework where a persistent reduction of 55 kcal/day results in a weight reduction of 0.4536 kg over one year, with half the change in the first twelve months. A dietary compensation parameter, γ, accounts for psychological licensing effects:

$$\Delta_{net} = \Delta_{kcal} \times (1 - \gamma)$$

Primary models assumed a central compensation parameter of 0.5.

Mortality modeling and survival engine

The microsimulation incorporates a survival engine to prevent immortal time bias. Baseline risk of death was established using ONS national life tables. Hazards $h_0(a, s)$ were dynamically adjusted in each cycle using BMI-specific risk factors. Mortality risk was adjusted log-linearly for BMI above 25 using a hazard ratio of 1.21 per 5-unit increase:

$$h_i(t) = h_0(a, s) \times 1.21^{\frac{BMI_i - 25}{5}}$$

Economic valuation and counterfactual scenarios

We evaluated consumer-only responses to coarse and detailed labels using transported PATEs, and combined these with mandatory 5 percent industry reformulation achieved

over 5 years. Reformulation costs were estimated based on annual sales. NHS healthcare burdens were estimated at £16 per excess BMI unit per person-year. Health utilities were modeled as a loss of 0.003 QALYs per BMI unit increase. Future values were discounted at 1.5 percent.

Model validation and sensitivity analyses

Internal validity was verified using common random numbers to isolate variance between arms. External validation compared simulated mortality against national data; the simulated 10-year all-cause mortality rate of 10.9 percent falls within observed confidence intervals. A probabilistic sensitivity analysis was executed using 1,000 Monte Carlo iterations to sample uncertainty in the causal intervention effect alongside health economic parameters. Dietary compensation rates were tested at 30, 50, and 70 percent.

Table C.7. Health gains and costs of coarse and detailed consumer FOP labelling policies under 30% dietary compensation

Model	Year	A. Coarse Label (Consumer Only)					B. Detailed Label (Consumer Only)				
		Kcal	Δ BMI	Obesity Prev.	QALYs	NHS (£m)	Kcal	Δ BMI	Obesity Prev.	QALYs	NHS (£m)
Baseline: Status Quo (Secular Trend)											
	Y1	-	0.01	-	-	£11.5	-	0.01	-	-	£11.5
	Y5	-	0.06	-	-	£136.0	-	0.06	-	-	£136.0
	Y10	-	0.11	-	-	£325.0	-	0.11	-	-	£325.0
Model A	Y1	-3.2	-0.03	79.8k	2.9k	-£15.4	2.4	0.02	-59.8k	-2.1k	£11.5
	Y5	-3.2	-0.05	118.5k	30.5k	-£146.2	2.4	0.03	-87.7k	-22.3k	£109.4
	Y10	-3.2	-0.05	99.0k	70.5k	-£296.1	2.4	0.03	-72.6k	-51.6k	£220.8
Model E	Y1	-4.8	-0.04	124.8k	4.4k	-£23.2	14.0	0.13	-366.6k	-12.8k	£66.8
	Y5	-4.8	-0.07	178.4k	47.9k	-£219.6	14.0	0.20	-525.7k	-143.4k	£632.3
	Y10	-4.8	-0.07	148.9k	114.2k	-£443.2	14.0	0.20	-444.0k	-337.1k	£1281.4
Model F	Y1	-3.0	-0.03	74.7k	2.7k	-£14.4	13.0	0.12	-340.9k	-11.9k	£63.0
	Y5	-3.0	-0.04	107.2k	28.6k	-£137.5	13.0	0.19	-487.4k	-135.2k	£598.2
	Y10	-3.0	-0.04	92.1k	66.0k	-£276.5	13.0	0.19	-410.5k	-317.0k	£1212.6

Table C.8. Health gains and costs of coarse and detailed FOP labelling combined with industry reformulation under 30% dietary compensation

Model	Year	C. Coarse + Reformulation					D. Detailed + Reformulation				
		Kcal	**Δ BMI**	**Obesity Prev.**	**QALYs**	**NHS (£m)**	**Kcal**	**Δ BMI**	**Obesity Prev.**	**QALYs**	**NHS (£m)**
Baseline: Status Quo (Secular Trend)											
	Y1	-	0.01	-	-	£11.5	-	0.01	-	-	£11.5
	Y5	-	0.06	-	-	£136.0	-	0.06	-	-	£136.0
	Y10	-	0.11	-	-	£325.0	-	0.11	-	-	£325.0
Model A	Y1	-4.9	-0.05	127.1k	4.5k	-£24.0	0.6	0.01	-14.5k	-554	£3.0
	Y5	-8.4	-0.16	404.0k	77.3k	-£357.5	-2.9	-0.08	199.5k	20.3k	-£101.9
	Y10	-12.8	-0.28	617.4k	267.2k	-£1097.0	-7.3	-0.20	444.2k	130.7k	-£574.7
Model E	Y1	-6.6	-0.06	171.1k	6.1k	-£31.7	12.2	0.11	-322.0k	-11.3k	£58.9
	Y5	-10.1	-0.18	469.4k	94.5k	-£431.2	8.7	0.09	-231.2k	-96.9k	£428.5
	Y10	-14.4	-0.31	659.5k	310.9k	-£1245.9	4.4	-0.04	80.5k	-139.4k	£496.3
Model F	Y1	-4.7	-0.04	120.4k	4.3k	-£22.7	11.3	0.10	-294.6k	-10.4k	£54.5
	Y5	-8.2	-0.16	395.7k	75.1k	-£345.2	7.8	0.08	-195.4k	-89.1k	£386.8
	Y10	-12.6	-0.28	607.4k	262.9k	-£1067.4	3.4	-0.05	113.5k	-119.0k	£410.0

References

[1] Ashkan Afshin et al. “Health effects of dietary risks in 195 countries, 1990–2017: a analysis for the Global Burden of Disease Study 2017”. In: *The lancet* 393.10184 (2019), pp. 1958–1972.

[2] Fernanda Rauber et al. “Ultra-processed food consumption and indicators of obesity in the United Kingdom population (2008-2016)”. In: *PloS one* 15.5 (2020), e0232676.

[3] Barry M Popkin et al. “Towards unified and impactful policies to reduce ultra-processed food consumption and promote healthier eating”. In: *The lancet Diabetes & endocrinology* 9.7 (2021), pp. 462–470.

[4] Ashkan Afshin et al. “Health effects of dietary risks in 195 countries, 1990–2017: a analysis for the Global Burden of Disease Study 2017”. In: *The lancet* 393.10184 (2019), pp. 1958–1972.

[5] Klaus G Grunert and Josephine M Wills. “A review of European research on consumer response to nutrition information on food labels”. In: *Journal of public health* 15 (2007), pp. 385–399.

[6] Zenobia Talati et al. “Consumersâ€™ perceptions of five front-of-package nutrition labels: An experimental study across 12 countries”. In: *Nutrients* 11.8 (2019), p. 1934.

[7] Iina Ikonen et al. “Consumer effects of front-of-package nutrition labeling: an interdisciplinary meta-analysis”. In: (2020). DOI: 10.1007/s11747-019-00663-9. URL: https://doi.org/10.1007/s11747-019-00663-9.

[8] Siyi Shangguan et al. “A Meta-Analysis of Food Labeling Effects on Consumer Diet Behaviors and Industry Practices”. In: *American journal of preventive medicine* 56 (2 Feb. 2019), pp. 300–314. ISSN: 1873-2607. DOI: 10.1016/J.AMEPRE.2018.09.024.

[9] Eleonora Fichera and Stephanie von Hinke. "The response to nutritional labels: Evidence from a quasi-experiment". In: *Journal of Health Economics* 72 (2020), p. 102326.

[10] Peggy J. Liu et al. "Using Behavioral Economics to Design More Effective Food Policies to Address Obesity". In: *Applied Economic Perspectives and Policy* 36 (1 Mar. 2014), pp. 6–24. ISSN: 2040-5790. DOI: 10.1093/aepp/ppt027.

[11] George Loewenstein, Cass R Sunstein, and Russell Golman. "Disclosure: Psychology changes everything". In: *Annu. Rev. Econ.* 6.1 (2014), pp. 391–419.

[12] Daniel Kahneman. "A Perspective on Judgment and Choice: Mapping Bounded Rationality". In: *American Psychologist* 58 (9 Sept. 2003), pp. 697–720. ISSN: 0003066X. DOI: 10.1037/0003-066X.58.9.697.

[13] Nano Barahona, Cristóbal Otero, and Sebastián Otero. "Equilibrium effects of food labeling policies". In: *Econometrica* 91.3 (2023), pp. 839–868.

[14] Edward Cartwright. *Behavioral economics*. London : Routledge, 2018.

[15] Lindsey Smith Taillie et al. "Experimental studies of front-of-package nutrient warning labels on sugar-sweetened beverages and ultra-processed foods: a scoping review". In: *Nutrients* 12.2 (2020), p. 569.

[16] Eric Andrew Finkelstein, Felicia Jia Ler Ang, and Brett Doble. "Randomized trial evaluating the effectiveness of within versus across-category front-of-package lower-calorie labelling on food demand". In: *BMC Public Health* 20 (2020), pp. 1–10.

[17] Silvio Ravaioli. "Coarse and Precise Information in Food Labeling". In: *Job Market Paper* (Oct. 2021). Times cited: 1, p. 1.

[18] Peter Scarborough et al. "Protocol for a pilot randomised controlled trial of an intervention to increase the use of traffic light food labelling in UK shoppers (the FLICC trial)". In: *Pilot and feasibility studies* 1 (2015), pp. 1–11.

[19] Rachel Griffith, Martin O'Connell, and Kate Smith. "The Importance of Product Reformulation Versus Consumer Choice in Improving Diet Quality". In: *Economica* 84 (333 2017). ISSN: 14680335. DOI: 10.1111/ecca.12192.

[20] Emily J Dhurandhar et al. "Predicting adult weight change in the real world: a systematic review and meta-analysis accounting for compensatory changes in energy intake or expenditure". In: *International journal of obesity* 39.8 (2015), pp. 1181–1187.

[21] Klaus G Grunert and Josephine M Wills. "A review of European research on consumer response to nutrition information on food labels". In: *Journal of public health* 15 (2007), pp. 385–399.

[22] Andrea Wilson. "Bounded Memory and Biases in Information Processing". In: *Econometrica* 82 (6 Nov. 2014), pp. 2257–2294. ISSN: 14680262. DOI: 10.3982/ECTA12188.

[23] Cecilia Lindig-León, Nehchal Kaur, and Daniel A Braun. "From Bayes-optimal to heuristic decision-making in a two-alternative forced choice task with an information-theoretic bounded rationality model". In: *Frontiers in Neuroscience* 16 (2022), p. 906198.

[24] Lorenzo M Donini et al. "Front-of-pack labels:â€œDirectiveâ€ versus â€œinformativeâ€ approaches". In: *Nutrition* 105 (2023), p. 111861.

[25] Michael Siegrist, Rebecca Leins-Hess, and Carmen Keller. "Which front-of-pack nutrition label is the most efficient one? The results of an eye-tracker study". In: *Food Quality and Preference* 39 (Oct. 2015). Times cited: 1, pp. 183–190. ISSN: 0950-3293. DOI: 10.1016/j.foodqual.2014.07.010.

[26] Food Standards Agency. *Guide to creating a front of pack (FoP) nutrition label for pre-packed products sold through retail outlets*. 2016.

[27] Paul Ayres and Fred Paas. *Cognitive load theory: New directions and challenges*. 2012.

[28] Andrew W Brown et al. “Toward more rigorous and informative nutritional epidemiology: the rational space between dismissal and defense of the status quo”. In: *Critical reviews in food science and nutrition* 63.18 (2023), pp. 3150–3167.

[29] Yu-Han Chiu et al. “Estimating the effect of nutritional interventions using observational data: the American Heart Associationâ€™s 2020 Dietary Goals and mortality”. In: *The American journal of clinical nutrition* 114.2 (2021), pp. 690–703.

[30] Yu-Han Chiu, Jorge E Chavarro, et al. “Well-defined interventions for nutritional studies: from target trials to nutritional modeling”. In: *Current Epidemiology Reports* 8 (2021), pp. 83–90.

[31] Bénédicte Colnet et al. “Causal inference methods for combining randomized trials and observational studies: a review”. In: *Statistical science* 39.1 (2024), pp. 165–191.

[32] Issa J Dahabreh et al. “Extending inferences from a randomized trial to a new target population”. In: *Statistics in medicine* 39.14 (2020), pp. 1999–2014.

[33] Pierre Dubois et al. “Effects of front-of-pack labels on the nutritional quality of supermarket food purchases: evidence from a large-scale randomized controlled trial”. In: *Journal of the Academy of Marketing Science volume* (2021). DOI: 10.1007/s11747-020-00723-5/Published. URL: https://doi.org/10.1007/s11747-020-00723-5.

[34] Paolo Crosetto et al. “Nutritional and economic impact of five alternative front-of-pack nutritional labels: experimental evidence”. In: (2020). DOI: 10.1093/erae/jbz037. URL: https://academic.oup.com/erae/article/47/2/785/5552528.

[35] Manon Egnell et al. “Objective understanding of Nutri-Score Front-Of-Package nutrition label according to individual characteristics of subjects: Comparisons with other format labels”. In: *PloS one* 13.8 (2018), e0202095.

[36] Paolo Crosetto, Laurent Muller, and Bernard Ruffieux. “Helping consumers with a front-of-pack label: Numbers or colors?: Experimental comparison between Guideline Daily Amount and Traffic Light in a diet-building exercise”. In: *Journal of Economic*

Psychology 55 (Aug. 2016), pp. 30–50. ISSN: 0167-4870. DOI: 10.1016/J.JOEP.2016.03.006.

[37] Marco Francesco Mazzù, Simona Romani, and Antea Gambicorti. "Effects on consumersâ€™ subjective understanding of a new front-of-pack nutritional label: a study on Italian consumers". In: *International journal of food sciences and nutrition* 72.3 (2021), pp. 357–366.

[38] Jessica Packer et al. "Secondary Outcomes of a Front-of-Pack-Labelling Randomised Controlled Experiment in a Representative British Sample: Understanding, Ranking Speed and Perceptions". In: *Nutrients* 14.11 (2022), p. 2188.

[39] Anne M Turner et al. "Recruiting older adult participants through crowdsourcing platforms: Mechanical Turk versus Prolific Academic". In: *AMIA Annual Symposium Proceedings*. Vol. 2020. 2021, p. 1230.

[40] Guido W Imbens. "Causal inference in the social sciences". In: *Annual Review of Statistics and Its Application* 11 (2024).

[41] Elizabeth A Stuart and Anna Rhodes. "Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data". In: *Evaluation review* 41.4 (2017), pp. 357–388.

[42] Maartje P Poelman et al. "Towards the measurement of food literacy with respect to healthy eating: the development and validation of the self perceived food literacy scale among an adult sample in the Netherlands". In: *International Journal of Behavioral Nutrition and Physical Activity* 15 (1 2018), p. 54. ISSN: 1479-5868. DOI: 10.1186/s12966-018-0687-z.

[43] Benjamin Ackerman et al. "Generalizing randomized trial findings to a target population using complex survey population data". In: *Statistics in medicine* 40.5 (2021), pp. 1101–1120.

[44] Bénédicte Colnet et al. “Re-weighting the randomized controlled trial for generalization: finite-sample error and variable selection”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 188.2 (2025), pp. 345–372.

[45] Catherine R Lesko et al. “The effect of antiretroviral therapy on all-cause mortality, generalized to persons diagnosed with HIV in the USA, 2009-11”. In: *International Journal of Epidemiology* 45.1 (2016), pp. 140–150.

[46] Megan S Schuler and Sherri Rose. “Targeted maximum likelihood estimation for causal inference in observational studies”. In: *American journal of epidemiology* 185.1 (2017), pp. 65–73.

[47] Bret Zeldow, Vincent Lo Re III, and Jason Roy. “A semiparametric modeling approach using Bayesian additive regression trees with an application to evaluate heterogeneous treatment effects”. In: *The annals of applied statistics* 13.3 (2019), p. 1989.

[48] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. “BART: Bayesian additive regression trees”. In: *The Annals of Applied Statistics* 4.1 (2010), pp. 266–298. DOI: 10.1214/09-AOAS285.

[49] Estevão B Prado, Rafael A Moral, and Andrew C Parnell. “Bayesian additive regression trees with model trees”. In: *Statistics and Computing* 31.3 (2021), p. 20.

[50] Falco J Bargagli-Stoffi et al. “Transporting Predictions via Double Machine Learning: Predicting Partially Unobserved Students’ Outcomes”. In: *arXiv preprint arXiv:2509.12533* (2025).

[51] Benjamin Ackerman et al. “Implementing statistical methods for generalizing randomized trial findings to a target population”. In: *Addictive behaviors* 94 (2019), pp. 124–132.

[52] Melody Huang and Samuel D Pimentel. “Variance-based sensitivity analysis for weighting estimators result in more informative bounds”. In: *arXiv preprint arXiv:2208.08381* (2022).

[53] Lan Nguyen and Hans De Steur. "Public Acceptability of Policy interventions to reduce sugary drink consumption in Urban Vietnam". In: *Sustainability* 13.23 (2021), p. 13422.

[54] Bénédicte Colnet et al. "Causal inference methods for combining randomized trials and observational studies: a review". In: *arXiv preprint arXiv:2003.04168* (2022).

[55] Fernanda Rauber et al. "Ultra-processed food consumption and risk of obesity: a prospective cohort study of UK Biobank". In: *European journal of nutrition* 60.4 (2021), pp. 2169–2180.

[56] Fernanda Rauber et al. "Ultra-processed food consumption and indicators of obesity in the United Kingdom population (2008-2016)". In: *PloS one* 15.5 (2020), e0232676.

[57] Montserrat Costa-Font, Wisdom Dogbe, and Cesar Revoredo-Giha. "Ready Meals in the UK: An Analysis Based on Their Nutritional and Sustainable Claims". In: (2021).

[58] Ala'a Alkerwi, Georgina E Crichton, and James R Hébert. "Consumption of ready-made meals and increased risk of obesity: findings from the Observation of Cardiovascular Risk Factors in Luxembourg (ORISCAV-LUX) study". In: *British Journal of Nutrition* 113.2 (2015), pp. 270–277.

[59] Véronique Braesco et al. "Ultra-processed foods: how functional is the NOVA system?" In: *European Journal of Clinical Nutrition* 76.9 (2022), pp. 1245–1253.

[60] SE Hillier, O Nunn, and K Lorrain-Smith. "An analysis of the nutritional value of UK supermarket ready meals". In: *Proceedings of the Nutrition Society* 79.OCE3 (2020), E794.

[61] Samuel J Dicken, Rachel L Batterham, and Adrian Brown. "Nutrients or processing? An analysis of food and drink items from the UK National Diet and Nutrition Survey based on nutrient content, the NOVA classification and front of package traffic light labelling". In: *British Journal of Nutrition* 131.9 (2024), pp. 1619–1632.

[62] Migena Luli et al. “The implications of defining obesity as a disease: a report from the Association for the Study of Obesity 2021 annual conference”. In: *EClinicalMedicine* 58 (2023).

[63] European Commission. *The Food and Beverage Market Entry Handbook: UNITED KINGDOM. A Practical Guide to the Market in the UK for European Agri-food Products*. Market Entry Handbook. Accessed August 1, 2025. European Commission, 2021.

[64] Michael A Rodriguez. “What makes a warning label salient?” In: *Proceedings of the human factors society annual meeting*. Vol. 35. 15. SAGE Publications Sage CA: Los Angeles, CA. 1991, pp. 1029–1033.

[65] Nicholas Mcinnes and Bo JA Haglund. “Readability of online health information: implications for health literacy”. In: *Informatics for health and social care* 36.4 (2011), pp. 173–189.

[66] Helen Croker et al. “Front of pack nutritional labelling schemes: a systematic review and meta-analysis of recent evidence relating to objectively measured consumption and purchasing”. In: *Journal of Human Nutrition and Dietetics* 33.4 (2020), pp. 518–537.

[67] Stewart Martin. “Measuring cognitive load and cognition: metrics for technology-enhanced learning”. In: *Technology-Enhanced and Collaborative Learning*. Routledge, 2018, pp. 77–106.

[68] Edward C Green, Elaine M Murphy, and Kristina Gryboski. “The health belief model”. In: *The Wiley encyclopedia of health psychology* (2020), pp. 211–214.

[69] Peter Scarborough et al. “Reds are more important than greens: how UK supermarket shoppers use the different information on a traffic light nutrition label in a choice experiment”. In: *International Journal of Behavioral Nutrition and Physical Activity* 12 (2015), pp. 1–9.

[70] Coady Wing, Kosali Simon, and Ricardo A Bello-Gomez. “Designing difference in difference studies: best practices for public health policy research”. In: *Annual review of public health* 39 (2018), pp. 453–469.

[71] Chris LS Coryn and Kristin A Hobson. “Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice”. In: *New Directions for Evaluation* 2011.131 (2011), pp. 31–39.

[72] Bradley H Wagenaar et al. “Using routine health information systems for well-designed health evaluations in low-and middle-income countries”. In: *Health policy and planning* 31.1 (2016), pp. 129–135.

[73] Marin L Schweizer, Barbara I Braun, and Aaron M Milstone. “Research methods in healthcare epidemiology and antimicrobial stewardshipâ€”quasi-experimental designs”. In: *Infection control & hospital epidemiology* 37.10 (2016), pp. 1135–1140.

[74] Paul R Hunter et al. “Impact of non-pharmaceutical interventions against COVID-19 in Europe in 2020: a quasi-experimental non-equivalent group and time series design study”. In: *Eurosurveillance* 26.28 (2021), p. 2001401.

[75] Jemma Hudson, Shona Fielding, and Craig R Ramsay. “Methodology and reporting characteristics of studies using interrupted time series design in healthcare”. In: *BMC medical research methodology* 19.1 (2019), p. 137.

[76] Ivlabèhiré Bertrand Meda, Seni Kouanda, and Valéry Ridde. “Effect of cost-reduction interventions on facility-based deliveries in Burkina Faso: a controlled interrupted time-series study with multiple non-equivalent dependent variables”. In: *J Epidemiol Community Health* 77.3 (2023), pp. 133–139.

[77] Sheryl L Chatfield et al. “Pre-test data and lessons learned from a group research project examining changes in physical activity behavior following construction of a rails-to-trails facility”. In: *Journal of community health* 39.2 (2014), pp. 386–393.

[78] Joshua J Reynolds. “Improving the assessment of teaching effectiveness with the nonequivalent dependent variables approach”. In: *Teaching of Psychology* 49.4 (2022), pp. 381–387.

[79] Gary Sacks, Mike Rayner, and Boyd Swinburn. “Impact of front-of-pack â€˜traffic-lightâ€™nutrition labelling on consumer food purchases in the UK”. In: *Health promotion international* 24.4 (2009), pp. 344–352.

[80] Jack P Hughes, Mario Weick, and Milica Vasiljevic. “Impact of pictorial warning labels on meat meal selection: A randomised experimental study with UK meat consumers”. In: *Appetite* 190 (2023), p. 107026.

[81] José Luis González-Castro et al. “Perceived vulnerability and severity predict adherence to COVID-19 protection measures: the mediating role of instrumental coping”. In: *Frontiers in Psychology* 12 (2021), p. 674032.

[82] Yin Wang, Jiayou Wang, and Qiong Shen. “A consumer segmentation study of nutrition information seeking and its relation to food consumption in Beijing, China”. In: *Foods* 11.3 (2022), p. 453.

[83] Food Standards Agency and National Centre for Social Research (NatCen). *Food and You, Waves 1-5 Data, 2010-2018*. Food Standards Agency. [data file]. 2019. URL: `https://www.food.gov.uk/research/food-and-you` (visited on 08/12/2025).

[84] Magaly Aceves-Martins et al. “Nutritional quality, environmental impact and cost of ultra-processed foods: a UK food-based analysis”. In: *International journal of environmental research and public health* 19.6 (2022), p. 3191.

[85] Angela M Odoms-Young et al. “Evaluating the initial impact of the revised Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) food packages on dietary intake and home food availability in African-American and Hispanic families”. In: *Public health nutrition* 17.1 (2014), pp. 83–93.

[86] Miriam Capasso et al. “Only the best for my kids: An extended TPB model to understand mothersâ€™ use of food labels”. In: *Appetite* 191 (2023), p. 107040.

[87] Siân Robinson et al. “Dietary patterns in infancy: the importance of maternal and family influences on feeding practice”. In: *British Journal of Nutrition* 98.5 (2007), pp. 1029–1037.

[88] Lindsey P Smith, Shu Wen Ng, and Barry M Popkin. “Trends in US home food preparation and consumption: analysis of national nutrition surveys and time use studies from 1965–1966 to 2007–2008”. In: *Nutrition journal* 12 (2013), pp. 1–10.

[89] Franziska Koch, Ingrid Hoffmann, and Erika Claupein. “Types of nutrition knowledge, their socio-demographic determinants and their association with food consumption: results of the NEMONIT study”. In: *Frontiers in nutrition* 8 (2021), p. 630014.

[90] Sara E Benjamin Neelon et al. “Spatial analysis of food insecurity and obesity by area-level deprivation in children in early years settings in England”. In: *Spatial and Spatio-temporal Epidemiology* 23 (Nov. 2017), pp. 1–9. ISSN: 18775845. DOI: 10.1016/j.sste.2017.07.001.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[92] Albee Y Ling et al. “An overview of current methods for real-world applications to generalize or transport clinical trial findings to target populations of interest”. In: *Epidemiology* 34.5 (2023), pp. 627–636.

[93] Daniel Celnik, Laura Gillespie, and MEJ Lean. “Time-scarcity, ready-meals, ill-health and the obesity epidemic”. In: *Trends in Food Science & Technology* 27.1 (2012), pp. 4–11.

[94] Alejandra Arrúa et al. “Warnings as a directive front-of-pack nutrition labelling scheme: comparison with the Guideline Daily Amount and traffic-light systems”. In: *Public health nutrition* 20.13 (2017), pp. 2308–2317.

[95] Laura Bix et al. “To See or Not to See: Do Front of Pack Nutrition Labels Affect Attention to Overall Nutrition Information?” In: *PLOS ONE* 10 (10 Oct. 2015), e0139732. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0139732. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139732.

[96] Soyun Kim and Michael S Wogalter. “Habituation, dishabituation, and recovery effects in visual warnings”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 53. 20. SAGE Publications Sage CA: Los Angeles, CA. 2009, pp. 1612–1616.

[97] Simon Howard, Jean Adams, and Martin White. “Nutritional content of supermarket ready meals and recipes by television chefs in the United Kingdom: cross sectional study”. In: *Bmj* 345 (2012).

[98] Rachel Griffith, Martin O’Connell, and Kate Smith. “Shopping around: how households adjusted food spending over the great recession”. In: *Economica* 83.330 (2016), pp. 247–280.

[99] Qëndresa Rramani et al. “Salient nutrition labels shift peoples’ attention to healthy foods and exert more influence on their choices”. In: *Nutrition Research* 80 (2020), pp. 106–116.

[100] Kathryn Cooper et al. “Exploring the Readability of Ingredients Lists of Food Labels with Existing Metrics”. In: *AMIA Summits on Translational Science Proceedings* 2022 (2022), p. 159.

[101] Scott Crossley et al. “A large-scaled corpus for assessing text readability”. In: *Behavior Research Methods* 55.2 (2023), pp. 491–507.

[102] Sebastian Araya et al. “Identifying Food Labeling Effects on Consumer Behavior”. In: *SSRN Electronic Journal* (Nov. 2019). DOI: 10.2139/SSRN.3195500. URL: https://papers.ssrn.com/abstract=3195500.

[103] Corinna Hawkes. “Nutrition labels and health claims: the global regulatory environment”. In: *Nutrition labels and health claims: the global regulatory environment.* 2004, pp. x–74.

[104] Peter Helfer and Thomas R Shultz. “The effects of nutrition labeling on consumer food choice: a psychological experiment and computational model”. In: *Annals of the New York Academy of Sciences* 1331.1 (2014), pp. 174–185.

[105] Olivier Allais, Fabrice EtilÃ©, and SÃ©bastien Lecocq. “Mandatory labels, taxes and market forces: An empirical evaluation of fat policies”. In: *Journal of Health Economics* 43 (Sept. 2015), pp. 27–44. ISSN: 0167-6296. DOI: 10.1016/J.JHEALECO.2015.06.003.

[106] Lindsey Smith Taillie et al. “An evaluation of Chile’s Law of Food Labeling and Advertising on sugar-sweetened beverage purchases from 2015 to 2017: A before-and-after study”. In: *PLoS medicine* 17 (2 Oct. 2020). Times cited: 1 PMID: 32045424 PMCID: PMC7012389, e1003015. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1003015.

[107] Eva-Maria Schruff-Lim et al. “Turning FOP nutrition labels into action: A systematic review of label+ interventions”. In: *Food Policy* 120 (2023), p. 102479.

[108] Valentina Pini et al. “Augmented grocery shopping: Fostering healthier food purchases through AR”. In: *Virtual Reality* 27.3 (2023), pp. 2117–2128.

[109] Simone Pettigrew et al. “The ability of nutrition warning labels to improve understanding and choice outcomes among consumers demonstrating preferences for unhealthy foods”. In: *Journal of the Academy of Nutrition and Dietetics* 124.1 (2024), pp. 58–64.

[110] Zhiyi Guo, Yueyue Ning, and Muhizam Mustafa. “Impact of Five Types of Front-of-Package Nutrition Labels on Consumer Behavior among Young Adults: A Systematic Review”. In: *Nutrients* 16.17 (2024), p. 2819.

[111] Shweta S Bapat, Harshali K Patel, and Sujit S Sansgiry. “Role of information anxiety and information load on processing of prescription drug information leaflets”. In: *Pharmacy* 5.4 (2017), p. 57.

[112] Wei Ji Ma, Masud Husain, and Paul M Bays. “Changing concepts of working memory”. In: *Nature neuroscience* 17.3 (2014), pp. 347–356.

[113] Roberto Dellâ€™Acqua et al. “On the functional independence of numerical acuity and visual working memory”. In: *Frontiers in Psychology* 15 (2024), p. 1335857.

[114] Neil A Lewis Jr and Allison Earl. “Seeing more and eating less: Effects of portion size granularity on the perception and regulation of food consumption.” In: *Journal of Personality and Social Psychology* 114.5 (2018), p. 786.

[115] Food Standards Agency. *Guide to creating a front of pack (FoP) nutrition label for pre-packed products sold through retail outlets*. 2016.

[116] Gijs A Holleman et al. “The â€˜real-world approachâ€™and its problems: A critique of the term ecological validity”. In: *Frontiers in Psychology* 11 (2020), p. 721.

[117] Meenakshie Bradley-Garcia and Victoria Bolton. “Programming an n-Back task in Qualtrics using HTML and JavaScript”. In: *University of Ottawa* 19 (2023).

[118] Rostam Golmohammadi et al. “Attention and short-term memory during occupational noise exposure considering task difficulty”. In: *Applied Acoustics* 158 (2020), p. 107065.

[119] Heike Schmidt et al. “No gender differences in brain activation during the N-back task: An fMRI study in healthy individuals”. In: *Human brain mapping* 30.11 (2009), pp. 3609–3615.

[120] Susanne M Jaeggi et al. “The concurrent validity of the N-back task as a working memory measure”. In: *Memory* 18.4 (2010), pp. 394–412.

[121] Eunice Jun, Gary Hsieh, and Katharina Reinecke. "Types of motivation affect study selection, attention, and dropouts in online experiments". In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), pp. 1–15.

[122] Harold Stanislaw and Natasha Todorov. "Calculation of signal detection theory measures". In: *Behavior research methods, instruments, & computers* 31.1 (1999), pp. 137–149.

[123] Simone Pettigrew et al. "The role of colour and summary indicators in influencing front-of-pack food label effectiveness across seven countries". In: *Public Health Nutrition* 24.11 (2021), pp. 3566–3570.

[124] Fumiaki Imamura et al. "Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction". In: *Bmj* 351 (2015).

[125] Scientific Advisory Committee on Nutrition. *Carbohydrates and Health*. Tech. rep. ISBN 978-0-11-708284-7. London: Public Health England, 2015. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/445503/SACN_Carbohydrates_and_Health.pdf.

[126] Rachel A Crockett et al. "Nutritional labelling for healthier food or non-alcoholic drink purchasing and consumption". In: *The Cochrane database of systematic reviews* 2018.2 (2018).

[127] Ana Laura Velazquez et al. "The influence of label information on the snacks parents choose for their children: Individual differences in a choice based conjoint test". In: *Food Quality and Preference* 94 (2021), p. 104296.

[128] Vincent Delhomme. "Front-of-pack nutrition labelling in the European Union: a behavioural, legal and political analysis". In: (2021). DOI: 10.1017/err.2021.5. URL: https://doi.org/10.1017/err.2021.5.

[129] Bridget Kelly and Jo Jewell. *What is the evidence on the policy specifications, development processes and effectiveness of existing front-of-pack food labelling policies in the WHO European Region?* World Health Organization. Regional Office for Europe, 2018.

[130] Lisa Zarantonello et al. "The effect of age, educational level, gender and cognitive reserve on visuospatial working memory performance across adult life span". In: *Aging, Neuropsychology, and Cognition* 27.2 (2020), pp. 302–319.

[131] World Health Organization. *Obesity and Overweight*. WHO Fact Sheet. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (visited on 01/15/2025).

[132] Marco Francesco Mazzù, Angelo Baccelloni, and Piera Finistauri. "Uncovering the effect of European policy-making initiatives in addressing nutrition-related issues: a systematic literature review and bibliometric analysis on front-of-pack labels". In: *Nutrients* 14.16 (2022), p. 3423.

[133] Michele Cecchini and Laura Warin. "Impact of food labelling systems on food choices and eating behaviours: a systematic review and meta-analysis of randomized studies". In: *Obesity reviews* 17.3 (2016), pp. 201–210.

[134] B. Ackerman et al. "Generalizing observational findings to a target population using propensity score methods: A note on survey weights". In: *Epidemiology* 31.5 (2020), pp. 739–742. DOI: 10.1097/EDE.0000000000001229.

[135] C. S. Buchanan, F. Li, and E. A. Stuart. "Generalizing treatment effects to a target population: a comparative study of weighting methods". In: *Statistical Methods in Medical Research* 27.10 (2018), pp. 3076–3094. DOI: 10.1177/0962280217739504.

[136] Quang Vuong et al. "Systematic review of applied transportability and generalizability analyses: A landscape analysis". In: *Annals of Epidemiology* (2025).

[137] Issa J Dahabreh et al. “Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population”. In: *Epidemiology* 31.3 (2020), pp. 334–344.

[138] F. Li et al. “Generalizing treatment effect estimates from randomized controlled trials to target populations”. In: *Statistical Science* 38.2 (2023), pp. 327–345. DOI: 10.1214/22-STS832.

[139] Ian Schmid et al. “Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations”. In: *Communications in Statistics-Simulation and Computation* 51.8 (2022), pp. 4326–4348.

[140] Naoki Egami and Erin Hartman. “Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 184.4 (2021), pp. 1524–1548.

[141] Donna L Coffman, Jiangxiu Zhou, and Xizhen Cai. “Comparison of methods for handling covariate missingness in propensity score estimation with a binary exposure”. In: *BMC medical research methodology* 20.1 (2020), p. 168.

[142] Trang Quynh Nguyen et al. “Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details”. In: *PloS one* 13.12 (2018), e0208795.

[143] Melody Y Huang. “Sensitivity analysis for the generalization of experimental results”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 187.4 (2024), pp. 900–918.

[144] Sanjay Basu et al. “Comparison of fruit and vegetable intake among urban low-income US adults receiving a produce voucher in 2 cities”. In: *JAMA network open* 4.3 (2021), e211757–e211757.

[145] Sarah E Robertson et al. “Comparing lung cancer screening strategies in a nationally representative US population using transportability methods for the National Lung Cancer Screening Trial”. In: *JAMA Network Open* 7.1 (2024), e2346295–e2346295.

[146] Pierre Chandon and Brian Wansink. “The biasing health halos of fast-food restaurant health claims: Lower calorie estimates and higher side-dish consumption intentions”. In: *Journal of Consumer Research* 34.3 (2007), pp. 301–314.

[147] Uzma Khan and Ravi Dhar. “Licensing effect in consumer choice”. In: *Journal of Marketing Research* 43.2 (2006), pp. 259–266.

[148] Constanza Avalos. “Food label granularity and working memory: effects on food choice in a randomized controlled trial”. In: *Journal of Health, Population and Nutrition* 44.1 (2025), pp. 1–16.

[149] A. R. Linero. “Bayesian regression trees for high dimensional prediction and variable selection”. In: *Journal of the American Statistical Association* 113.521 (2018), pp. 364–374. DOI: 10.1080/01621459.2017.1309852.

[150] A. Kern et al. “Assessing the role of neighborhood effects in the effectiveness of housing mobility programs: a nonparametric Bayesian approach”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179.1 (2016), pp. 89–111. DOI: 10.1111/rssa.12130.

[151] Junxiu Liu et al. “Health and economic impacts of the National Menu Calorie Labeling Law in the United States: a microsimulation study”. In: *Circulation: Cardiovascular Quality and Outcomes* 13.6 (2020), e006313.

[152] Carolyn M Rutter, Alan M Zaslavsky, and Eric J Feuer. “Dynamic microsimulation models for health outcomes: a review”. In: *Medical Decision Making* 31.1 (2011), pp. 10–18.

[153] Nadia Flexner et al. “Estimating the dietary and health impact of implementing front-of-pack nutrition labeling in Canada: A macrosimulation modeling study”. In: *Frontiers in Nutrition* 10 (2023), p. 1098231.

[154] Georgia D Tomova et al. “Adjustment for energy intake in nutritional research: a causal inference perspective”. In: *The American journal of clinical nutrition* 115.1 (2022), pp. 189–198.

[155] Fernando Alarid-Escudero et al. “Requirements for Data Reporting and Validation of Microsimulation Models: A Review”. In: *Medical Decision Making* 40.5 (2020), pp. 626–639.

[156] Public Health England., and Food Standards Agency. *National Diet and Nutrition Survey Rolling Programme, Years 1-15, 2008-2023*. [data collection]. UK Data Service. SN: 6533. 2024. DOI: 10.5255/UKDA-SN-6533-18.

[157] Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67. DOI: 10.18637/jss.v045.i03.

[158] Fan Li, Ashley L Buchanan, and Stephen R Cole. “Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 71.3 (2022), pp. 669–697.

[159] Linh Tran et al. “Double robust efficient estimators of longitudinal treatment effects: comparative performance in simulations and a case study”. In: *The international journal of biostatistics* 15.2 (2019), p. 20170054.

[160] Mikel Hernandez et al. “Synthetic data generation for tabular health records: A systematic review”. In: *Neurocomputing* 493 (2022), pp. 28–45.

[161] Lisa Pilgram et al. “Should we synthesize more than we need: impact of synthetic data generation for high-dimensional cross-sectional medical data”. In: *Journal of the American Medical Informatics Association* 32.12 (2025), pp. 1843–1854.

[162] A Kiran and S Saravana Kumar. “A methodology and an empirical analysis to determine the most suitable synthetic data generator”. In: *IEEE Access* 12 (2024), pp. 12209–12228.

[163] James Robards. “National population projections: 2020-based interim”. In: *Office for National Statistics* (2022).

[164] Kevin D Hall et al. “Quantification of the effect of energy imbalance on bodyweight”. In: *The Lancet* 378.9793 (2011), pp. 826–837.

[165] Public Health England. *Sugar reduction: achieving the 20%. A technical report outlining progress to date, guidelines for industry, 2015 baseline levels in key foods and next steps*. Tech. rep. Accessed March 20, 2020. London: Public Health England, 2017. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/604336/Sugar_reduction_achieving_the_20_.pdf.

[166] F. J. He, H. C. Brinsden, and G. A. MacGregor. “Salt reduction in the United Kingdom: a successful experiment in public health”. In: *Journal of Human Hypertension* 28.6 (2014), pp. 345–352. DOI: 10.1038/jhh.2013.105.

[167] Office for National Statistics. *Death registration summary statistics, England and Wales: 2024*. ONS website, statistical bulletin. Released 20 May 2025. May 2025. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathregistrationsummarystatisticsenglandandwales/2024.

[168] Krishnan Bhaskaran et al. “Association of BMI with overall and cause-specific mortality: a population-based cohort study of $3 \cdot 6$ million adults in the UK”. In: *The lancet Diabetes & endocrinology* 6.12 (2018), pp. 944–953.

[169] Heli Hytti, Reijo Takalo, and Heimo Ihalainen. “Tutorial on multivariate autoregressive modelling”. In: *Journal of clinical monitoring and computing* 20.2 (2006), pp. 101–108.

[170] Seamus Kent et al. “Quantifying the impact of obesity on NHS costs”. In: *Diabetes, Obesity and Metabolism* 19.9 (2017), pp. 1277–1285.

[171] Adam DM Briggs et al. “Estimating the cost-effectiveness of salt reformulation and increasing access to leisure centres in England, with PRIMEtime CE model validation using the AdViSHE tool”. In: *BMC Health Services Research* 19.1 (2019), pp. 1–13.

[172] Mark Freeman, Ben Groom, and Michael Spackman. “Social discount rates for cost-benefit analysis: a report for HM Treasury”. In: *HM Treasury: London, UK* (2018).

[173] Peggy J Liu et al. “Using behavioral economics to design more effective food policies to address obesity”. In: *Applied Economic Perspectives and Policy* 36.1 (2014), pp. 6–24.

[174] Meng Shen, Lijia Shi, and Zhifeng Gao. “Beyond the food label itself: How does color affect attention to information on food labels and preference for food attributes?” In: *Food Quality and Preference* 64 (Mar. 2018), pp. 47–55. ISSN: 0950-3293. DOI: 10.1016/J.FOODQUAL.2017.10.004.

[175] Ya Hui Michelle See, Richard E Petty, and Lisa M Evans. “The impact of perceived message complexity and need for cognition on information processing and attitudes”. In: *Journal of Research in Personality* 43.5 (2009), pp. 880–889.

[176] Constanza Avalos, Nick Shryane, and Yan Wang. “Food Label Readability and Consumption Frequency: Isolating Content-Specific Effects via a Non-Equivalent Dependent Variable Design”. In: *Nutrients* 18 (2026), p. 197.

[177] Constanza Avalos. “Food label granularity and working memory: effects on food choice in a randomized controlled trial”. In: *Journal of Health, Population and Nutrition* 44.375 (2025). DOI: 10.1186/s41043-025-01076-x.

[178] Clare Whitton et al. “National Diet and Nutrition Survey: UK food consumption and nutrient intakes from the first year of the rolling programme and comparisons with previous surveys”. In: *British journal of nutrition* 106.12 (2011), pp. 1899–1914.

[179] Arpita Saha. “Causal inference methods for policy evaluation”. PhD thesis. University of Manchester, 2022.

[180] Marcela Reyes et al. “Changes in the amount of nutrient of packaged foods and beverages after the initial implementation of the Chilean Law of Food Labelling and Advertising: A nonexperimental prospective study”. In: *PLoS medicine* 17.7 (2020), e1003220.

[181] Rachel Griffith, Martin O’Connell, and Kate Smith. “The importance of understanding the food industry for public health: a review of the evidence”. In: *Journal of Economic Literature* (2017).

[182] O Turnbull and H Ensaff. “An examination of food security and fruit and vegetable consumption using the Food and You Survey”. In: *Proceedings of the Nutrition Society* 77.OCE4 (2018), E190.

[183] Eleanor Hayes-Larson et al. “Estimating dementia incidence in insured older Asian Americans and Pacific Islanders in California: an application of inverse odds of selection weights”. In: *American journal of epidemiology* 194.5 (2025), pp. 1304–1313.

[184] Sreeram V Ramagopalan et al. “Transportability of overall survival estimates from US to Canadian patients with advanced non–small cell lung cancer with implications for regulatory and health technology assessment”. In: *JAMA Network Open* 5.11 (2022), e2239874–e2239874.

[185] Eva H DuGoff, Megan Schuler, and Elizabeth A Stuart. “Generalizing observational study results: applying propensity score methods to complex surveys”. In: *Health services research* 49.1 (2014), pp. 284–303.