



## Do we eat what we buy? Relative validity of grocery purchase data as an indicator of food consumption in the LoCard study

Henna Vepsäläinen<sup>1\*</sup>, Jaakko Nevalainen<sup>2</sup>, Satu Kinnunen<sup>1</sup>, Suvi Tuulikki Itkonen<sup>1</sup>, Jelena Meinilä<sup>1</sup>, Satu Männistö<sup>3</sup>, Liisa Uusitalo<sup>1</sup>, Mikael Fogelholm<sup>1</sup> and Maijaliisa Erkkola<sup>1</sup>

<sup>1</sup>Department of Food and Nutrition, University of Helsinki, PO Box 66, 00014 Helsinki, Finland

<sup>2</sup>Faculty of Social Sciences (Health Sciences), Tampere University, Arvo Ylpön katu 34, 33014 Tampere, Finland

<sup>3</sup>Finnish Institute for Health and Welfare (THL), PO Box 30, 00271 Helsinki, Finland

(Submitted 8 April 2021 – Final revision received 16 September 2021 – Accepted 13 October 2021 – First published online 18 October 2021)

### Abstract

The validity of grocery purchase data as an indicator of food consumption is uncertain. This paper investigated (1) the associations between food consumption and grocery purchases using automatically accumulated purchase data and (2) whether the strength of the associations differed in certain sub-populations. The participants filled in a FFQ, and a major Finnish retailer issued us with their loyalty-card holders' grocery purchase data covering the 1- and 12-month periods preceding the FFQ. We used gamma statistics to study the association between thirds/quarters of FFQ and grocery purchase data (frequency/amount) separately for eighteen food groups among the 11 983 participants. Stratified analyses were conducted for subgroups based on sex, family structure, educational level, household income and self-estimated share of purchases from the retailer. We also examined the proportion of participants classified into the same, adjacent, subsequent and opposite categories using the FFQ and purchase data. The gammas ranged from 0.12 (cooked vegetables) to 0.75 (margarines). Single households had stronger gammas than two-adult families, and participants with > 60 % of purchases from the retailer had stronger gammas. For most food groups, the proportion of participants classified into the same or adjacent category was > 70 %. Most discrepancies were observed for fresh/cooked vegetables, berries and vegetable oils. Even though the two methods did not categorise all food groups similarly, we conclude that grocery purchase data are able to describe food consumption in an adult population, and future studies should consider purchase data as a resource-saving and moderately valid measure in large samples.

**Key words:** Big data: Dietary assessment: Epidemiological study: Supermarkets: Validation study

In order to prevent non-communicable diseases, societies need to engage in evidence-based decision-making<sup>(1,2)</sup>. For this purpose, robust and accurate information on population-level health behaviours, such as diet, is essential. Most commonly, large, nation-wide surveys are used to monitor diet, but these methods entail a wide range of challenges such as difficulties in assessing diet<sup>(3–6)</sup> and declining participation rates<sup>(7,8)</sup>. The surveys aim for a representative sample, but in reality, certain population groups, such as women, highly educated and those with favourable health, tend to have higher response rates (selection bias)<sup>(7,9)</sup>. In addition, the aforementioned methods are slow to respond to, for example, rapid changes due to food policy measures (e.g. imposing or revising food-related taxes). As a result, surveys may under- or over-represent certain population groups and the data might already be outdated once the results are published. To overcome these problems, novel approaches to dietary assessment are constantly being

developed. One such alternative is the use of grocery purchase data as an indicator of food consumption.

Grocery purchase data collection can vary from home food inventories, which describe the availability of foods and beverages in the home at a single point in time, to collecting purchase receipts or recording all purchased foods and scanning food purchases in the store<sup>(10)</sup>. Compared with traditional dietary assessment methods (e.g. food records, 24-h dietary recalls), the collection of grocery purchase data, if at least partly automated, can be less burdensome for both participants and researchers and, in theory, could reach more than a million individuals or households<sup>(11)</sup>. However, it is not clear whether grocery purchase data can be used to assess individual-level food consumption, as the purchased foods may be consumed by other individuals and not all foods are purchased from one retailer<sup>(12)</sup>.

To date, a substantial body of literature concerning the concordance between grocery purchases and food consumption has

\* Corresponding author: Henna Vepsäläinen, email [henna.vepsalainen@helsinki.fi](mailto:henna.vepsalainen@helsinki.fi)

started to accumulate. However, the number of studies is still small and as diverse data collection methods have been applied, caution must be exercised when interpreting the results. Nevertheless, at least a moderate correlation seems to exist between grocery purchases and individual-level food consumption<sup>(13–19)</sup>, although weaker associations have also been reported<sup>(20)</sup>. The aforementioned studies have, however, required active participation, as grocery purchases were reported by the participants, which may have led to both misreporting and unrepresentativeness in the samples. In our recent study, automatically accumulating loyalty-card data on beer purchase frequency was found to associate with self-reported beer-drinking frequency<sup>(21)</sup>, but to the best of our knowledge, loyalty-card data have not been used to study the validity of grocery purchases as an indicator of food consumption.

In the current study, we examined (1) the associations between the participants' food consumption and grocery purchases using purchase data from 1- and 12-month periods and (2) whether the strength of the associations differed in certain sub-populations (e.g. men and women, families with and without children). To fulfil these objectives, we compared the food consumption frequencies and grocery purchases of 11 983 participants using an exceptional data set of automatically accumulated loyalty-card data on grocery purchases.

## Methods

### Study sample

In June 2018, S Group (S-Ryhmä, a major Finnish retail chain with a market share of 46 % in 2018<sup>(22)</sup>) loyalty-card holders received an email asking for their electronic informed consent to obtain their purchase data from 1 September 2016 to 31 December 2018. We did not contact card holders who had no declared email address, those prohibiting the retailer contacting them for any marketing or research-related issues and card holders under 18 years of age. Altogether 47 066 participants consented to participate in the study (see Fig. 1 for participant flow chart). A more detailed description of the study design and sample has been published earlier<sup>(23)</sup>. Before handover to the research team, the data were pseudonymised. The University of Helsinki Review Board in Humanities and Social and Behavioural Sciences reviewed the study and found it to be ethically acceptable (Statement 21/2018).

### Food consumption assessment

In June 2018, the consenting participants were randomly assigned to fill in one of the three electronic questionnaires developed for the study. Thus, a third ( $n$  15 756, 33 %) of them filled in a FFQ describing their food consumption during the past month. Of these participants, 11 983 (76 %) provided FFQ data. The FFQ was designed to measure the whole diet of the participants and was based on a previous FFQ used in the DAGIS study<sup>(24)</sup>. It included fifty-two food items and had six answer options: not at all, less than once a month, on 1–3 days in a month, on 1–2 days in a week, on 3–5 days in a week and daily or almost daily. These were further transformed into times/week

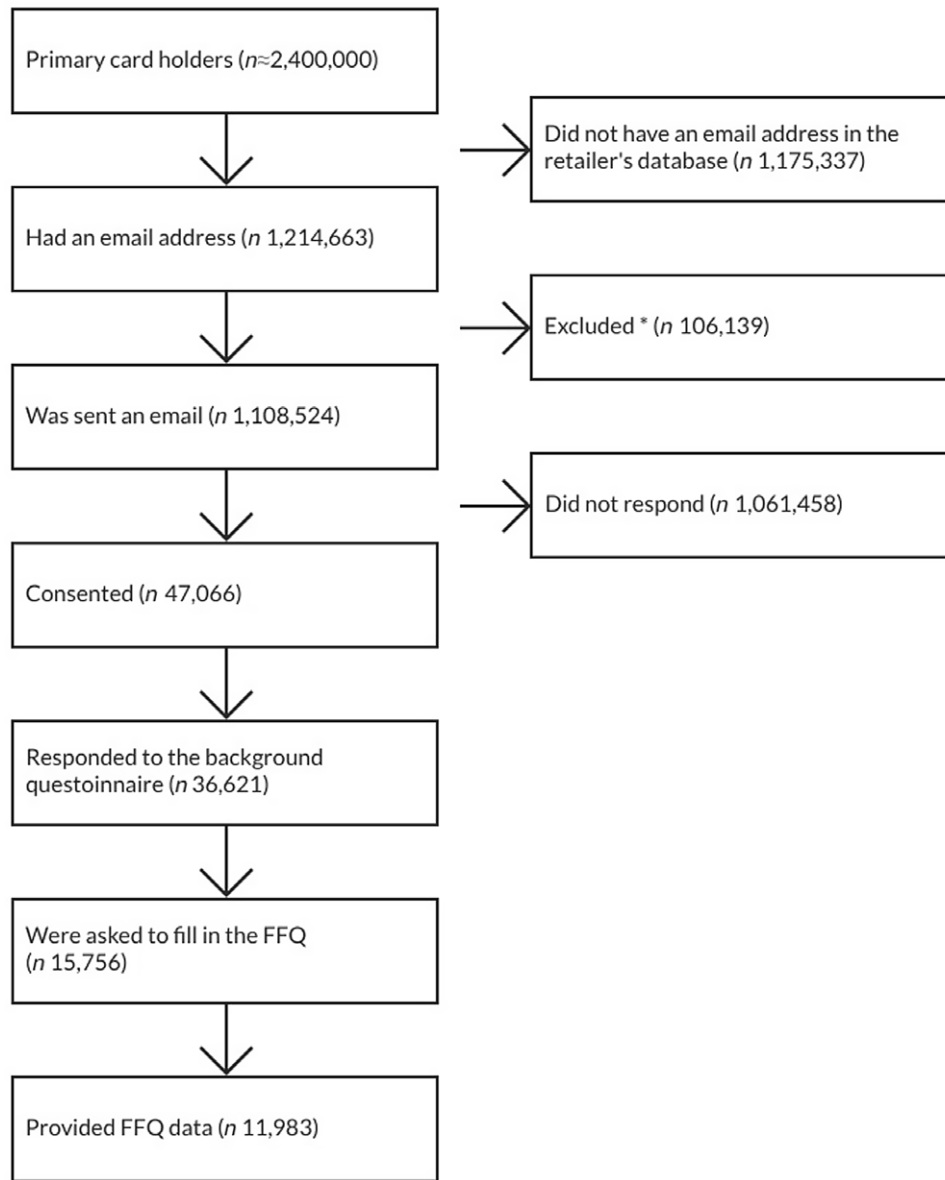
(0, 0–12, 0–47, 1–5, 4 and 6, respectively) to allow summation of consumption frequencies. We selected twenty-two FFQ rows for further inspection and collapsed them into eighteen food consumption variables describing the consumption frequencies of (1) fresh vegetables; (2) cooked and canned vegetables; (3) peas, beans, lentils and soya; (4) fruits; (5) berries; (6) skimmed milk and sour milk; (7) vegetable oils (used in cooking); (8) butter and butter-fat blends (on bread); (9) margarines (on bread); (10) fish dishes and fish products; (11) red meat; (12) white meat; (13) processed meats (cold cuts and sausages); (14) rye bread; (15) white bread; (16) sugar-sweetened beverages (juices and soft drinks); (17) chocolates and (18) sweets. The aforementioned foods were chosen on the basis of their role as indicators of a healthy diet. Online Supplementary Table 1 shows the consumption frequencies for the original FFQ rows.

### Grocery purchase data

The original purchase data consisted of 4234 grocery product groups, which had been created by the retailer. Of these, 3340 (79 %) were included in the food group classification created for the LoCard study (<https://locard.fi/inenglish/>). A professional nutritionist reclassified the grocery product groups received from the retailer into nutritionally meaningful food groups. Mixed dishes and rarely purchased product groups with no definite primary ingredient (altogether thirty-eight product groups) were left unclassified. The current study used the purchase data of 1301 product groups, which were assigned into thirty-nine food groups and further collapsed into eighteen variables to allow comparison with the FFQ. The variables describe the purchases of (1) fresh vegetables; (2) cooked and canned vegetables, mushrooms and vegetable dishes; (3) peas, beans, lentils, soya and plant-protein products; (4) fresh, canned and frozen fruit; (5) fresh and frozen berries; (6) skimmed milk and sour milk; (7) vegetable oils; (8) butter-fat blends; (9) margarines; (10) fish dishes, fish products and fresh fish; (11) cooked and fresh red meat, meat patties and balls, offal and red meat dishes; (12) cooked and fresh poultry, poultry offal, patties and dishes; (13) cold cuts, ham, bacon and sausages; (14) rye bread; (15) wheat bread; (16) sugar-sweetened juices, soft drinks and energy drinks; (17) chocolates and (18) sweets (see online Supplementary Table 2 for matching the grocery purchase categories and FFQ rows). In the analyses, we used purchase frequencies (number of days on which food items were purchased) and purchase volumes (in kg) from 1- and 12-month periods preceding the collection of the questionnaire data, including the FFQ data. Purchase frequencies and volumes are shown in online Supplementary Table 3, and Supplementary Fig. 1 differentiates between the FFQ and purchase data types used in the analyses.

### Background information

We obtained information on the participants' sex and age from the retailer's database. Using an electronic questionnaire, the participants reported how many members belonged to their household at that time. In addition, they were asked to separately report how many of the household members were



**Fig. 1.** The flow of participants in the current study. \* Reasons for exclusions: (1) under 18 year of age or (2) card holder had prohibited any contact with marketing or research-related material.

0–6-year-olds, 7–17-year-olds, 18–24-year-olds, 25–64-year-olds and 65-year-olds or older. Following closely the conventional family structure categories by Statistics Finland<sup>(25,26)</sup>, the participants were classified into five family structure categories: single-adult households, one adult and a child/children, two adults, two adults and a child/children, or other (households with three or more adults and households with unknown family structure).

Four categories (primary school or lower, upper secondary school, Bachelor's degree or equivalent, Master's degree or higher) were used to assess the participants' educational level. In addition, the participants reported their household monthly income using seven predefined options ranging from less than 1500–9000 €/month or more. To calculate scaled monthly household income, the mean income in each of the categories

was divided by the square root of household size<sup>(27)</sup>. Scaled monthly household income was then classified into five categories (less than 1000, 1000–1999, 2000–2999, 3000–3999 and 4000 €/month or more). The participants also estimated how large a proportion of their overall household's food purchases was bought from the retailer providing the purchase data (20 % or less, 21–40 %, 41–60 %, 61–80 % or more than 80 %).

*Statistical methods*

The FFQ data (food consumption frequencies in times/week) were divided into quarters using quartile classification. For certain food groups, the responses aggregated strongly, and thus, tertile classification was used for consumption frequencies of fresh vegetables, cooked and canned vegetables, legumes,

red meat, white meat, chocolate and sweets. Each of the grocery purchase variables was categorised to a comparable scale (e.g. 1–4 or 1–3) based on cut-points defined by the quartiles and tertiles of their distribution.

We used gamma statistics<sup>(28)</sup> with 95 % CI to study the relative validity of the purchase data, that is, the association between the ordinal categories (thirds or quarters) of the two sources of data (FFQ and purchase data)<sup>(3)</sup>. The gamma statistic measures the strength of association when both variables are ordinal and its value ranges from –1 to +1. Based on the cut-offs set by BMJ, we regarded gammas as very weak (0.19 or lower), weak (0.20–0.39), moderate (0.40–0.59), strong (0.60–0.79) or very strong (higher than 0.80)<sup>(29)</sup>. In order to examine internal consistency<sup>(30)</sup>, we compared the different types of grocery purchase data (purchase frequency and purchase volume during the past 1- and 12-month periods) with the FFQ data, separately for the eighteen food groups. In subgroup analyses, we calculated the gammas and 95 % CI comparing the FFQ and 12-month purchase data separately for (a) men and women, (b) single-adult households and two-adult families with no children, (c) one- and two-adult families with children, (d) participants with low (upper secondary school or less) and high (Bachelor's level or higher) education, (e) households with low (less than 1000 €) and high (4000 € or more) scaled monthly income and (f) participants with low (60 % or less) or high (more than 60 %) degree of loyalty. Complete case approach was used for the subgroup analyses. Furthermore, as participants were assigned to categories (thirds or quarters defined by quantile cut points) based on both food consumption frequencies and 12-month volume of grocery purchases, we calculated the proportion of participants classified into the same, adjacent (+/–1), subsequent (+/–2) or opposite categories to describe the agreement between the two methods. We investigated the distributions of the differences to assess whether the two methods could rank the participants in the same order. We used free statistical software R (R Foundation for Statistical Computing, <http://www.R-project.org/>) for the analyses.

**Results**

Altogether 11 983 participants (76 % of those randomised to fill in the FFQ) provided FFQ data and were included in the analyses. Their mean age was 48 years (SD 15 years, range 18–93 years) and two-thirds of them were women (Table 1). Approximately one-third of the participants lived with another adult and had no children in the household, whereas a quarter of the participants lived in a single-adult household. In addition, a quarter of the participants lived in a household with a child or children. Two-thirds of the participants (64 %) reported that they made more than 60 % of their grocery purchases from the data-providing retailer.

Table 2 describes the gammas for each of the food groups using the four different types of purchase data. Overall, the gammas ranged from 0.12 (cooked and canned vegetables) to 0.75 (margarines). The strongest associations were observed for skimmed milk and sour milk as well as for margarines: the gammas for these ranged between 0.63–0.74 and

**Table 1.** Descriptive characteristics of participants who provided FFQ data (numbers and percentages, *n* 11 983)

		<i>n</i>	%
Sex	Men	3999	33
	Women	7979	67
	Missing	5	0.04
Family structure	Single-adult household	3152	26
	One adult and a child/children	500	4.2
	Two adults, no children	4274	36
	Two adults and a child/children	2506	21
	Other*	900	7.5
Educational level	Missing	651	5.4
	Primary school or lower	731	6.1
	Upper secondary school	4367	36
	Bachelor's degree or equivalent	3922	33
	Master's degree or higher	2915	24
Proportion of purchases from retailer	Missing	48	0.40
	20 % or less	755	6.3
	21–40 %	1551	13
	41–60 %	2007	17
	61–80 %	3026	25
	More than 80 %	4630	39
	Missing	14	0.12
Weighted monthly household income, €†	Less than 1000	1003	8.4
	1000–1999	1720	14
	2000–2999	3431	29
	3000–3999	2714	23
	4000 or more	2227	19
	Missing	888	7.4

\* Households with three or more adults and households with unknown family structure.  
 † Mean household monthly income divided by square root of household size reported by participants.

0.63–0.75 depending on the type of grocery purchase data used. For butter and butter-fat blends, legumes and sweets, we detected mostly moderate gammas (0.48–0.53, 0.36–0.47 and 0.45–0.52, respectively), whereas for cooked and canned vegetables, the gammas were very weak (0.12–0.15). Different types of purchase data gave fairly similar associations. On average, the 12-month purchase volume showed slightly stronger associations with food consumption variables than other types of purchase data.

Figure 2 shows the comparison between different participant subgroups' gammas. For most of the food groups, the gammas of men and women did not differ. However, women had higher gammas for red meat, legumes and processed meats, whereas men's gammas were higher for fruits and vegetable oils. Not surprisingly, single-adult households had stronger gammas than two-adult families. However, the gammas of one- and two-adult families with children did not differ. Only a few differences were observed between levels of education, but for red meat, white meat and processed meats, participants with higher education had stronger gammas. Compared with the participants with a scaled household income of 4000 €/month or more, the participants with a scaled household income of 1000 €/month or less had higher gammas for fresh vegetables, legumes, fruits, and cooked and canned vegetables. For margarines, the gammas

**Table 2.** Associations between quarters and thirds (depending on distribution of variables) of food consumption frequencies and grocery purchase frequencies and volumes over past 1- and 12-month periods; numbers shown are gammas and 95 % confidence intervals, *n* 11 543–11 926. Both FFQ and purchase data were categorised into thirds unless otherwise indicated

	1-month period				12-month period			
	Purchase frequency		Purchase volume		Purchase frequency		Purchase volume	
	Gamma	Lower CI, upper CI	Gamma	Lower CI, upper CI	Gamma	Lower CI, upper CI	Gamma	Lower CI, upper CI
Fresh vegetables	0.33	0.30, 0.36	0.39	0.36, 0.41	0.34	0.32, 0.37	0.42	0.39, 0.44
Cooked and canned vegetables	0.13	0.11, 0.16	0.14	0.12, 0.17	0.12	0.1, 0.14	0.15	0.13, 0.18
Legumes	0.47	0.44, 0.5	0.46	0.43, 0.49	0.41	0.38, 0.43	0.36	0.34, 0.39
Fruits*	0.33	0.31, 0.34	0.38	0.36, 0.39	0.3	0.28, 0.32	0.37	0.35, 0.39
Berries†	0.30	0.28, 0.33	0.30	0.27, 0.33	0.26	0.24, 0.28	0.27	0.25, 0.29
Skimmed milk and sour milk‡	0.74	0.72, 0.76	0.74	0.73, 0.76	0.63	0.61, 0.64	0.63	0.62, 0.65
Vegetable oils‡	0.34	0.3, 0.37	0.34	0.30, 0.37	0.31	0.29, 0.33	0.29	0.27, 0.31
Butter and butter-fat blends‡	0.52	0.49, 0.54	0.48	0.47, 0.50	0.51	0.49, 0.52	0.53	0.51, 0.55
Margarines§	0.66	0.64, 0.68	0.75	0.73, 0.76	0.63	0.62, 0.65	0.63	0.62, 0.65
Fish*	0.30	0.28, 0.32	0.30	0.28, 0.33	0.33	0.31, 0.35	0.35	0.33, 0.37
Red meat	0.34	0.31, 0.36	0.37	0.34, 0.39	0.34	0.32, 0.36	0.39	0.37, 0.41
White meat	0.39	0.37, 0.41	0.39	0.37, 0.42	0.4	0.37, 0.42	0.41	0.39, 0.43
Processed meats*	0.36	0.34, 0.38	0.38	0.36, 0.40	0.38	0.36, 0.39	0.42	0.4, 0.44
Rye bread*	0.38	0.34, 0.38	0.37	0.35, 0.39	0.34	0.32, 0.36	0.37	0.35, 0.39
Wheat bread‡	0.24	0.22, 0.27	0.22	0.2, 0.25	0.25	0.23, 0.27	0.26	0.24, 0.28
Sugar-sweetened beverages‡	0.37	0.35, 0.39	0.35	0.33, 0.37	0.4	0.38, 0.41	0.4	0.38, 0.42
Chocolate	0.37	0.35, 0.4	0.36	0.34, 0.39	0.37	0.35, 0.4	0.35	0.33, 0.37
Sweets	0.45	0.43, 0.47	0.46	0.44, 0.48	0.48	0.46, 0.50	0.52	0.5, 0.54

\* Both FFQ and purchase data categorised into quarters.

† FFQ data categorised into thirds and quarters, thirds used for 1-month purchase data and quarters for 12-month purchase data.

‡ FFQ data categorised into thirds and quarters, thirds used for 1-month purchase frequency, quarters used for other purchase data.

§ FFQ data categorised into thirds and quarters, thirds used for 1-month purchase volume, quarters for other purchase data.

were stronger in the higher income group. Participants with more than 60 % of food purchases from the retailer had stronger gammas for all other food groups except for legumes, berries, and cooked and canned vegetables.

The proportion of participants classified into the same category using the FFQ and 12-month volume purchase data was 37–49 % for the food groups divided into thirds, and 31–49 % for the food groups divided into quarters (Fig. 3). Regarding sweets, red meat, skimmed milk and sour milk, margarines and processed meats, the two methods (FFQ data and purchase data) classified the participants relatively similarly (38–49 % of participants in the same category and 4–11 % in the opposite category). Largest discrepancies between the methods were observed for fresh vegetables, cooked vegetables, berries and vegetable oils (31–42 % of the participants in the same category, 12–21 % in the opposite category). For almost all food groups, the proportion of participants classified into the same or adjacent category was more than 70 %.

## Discussion

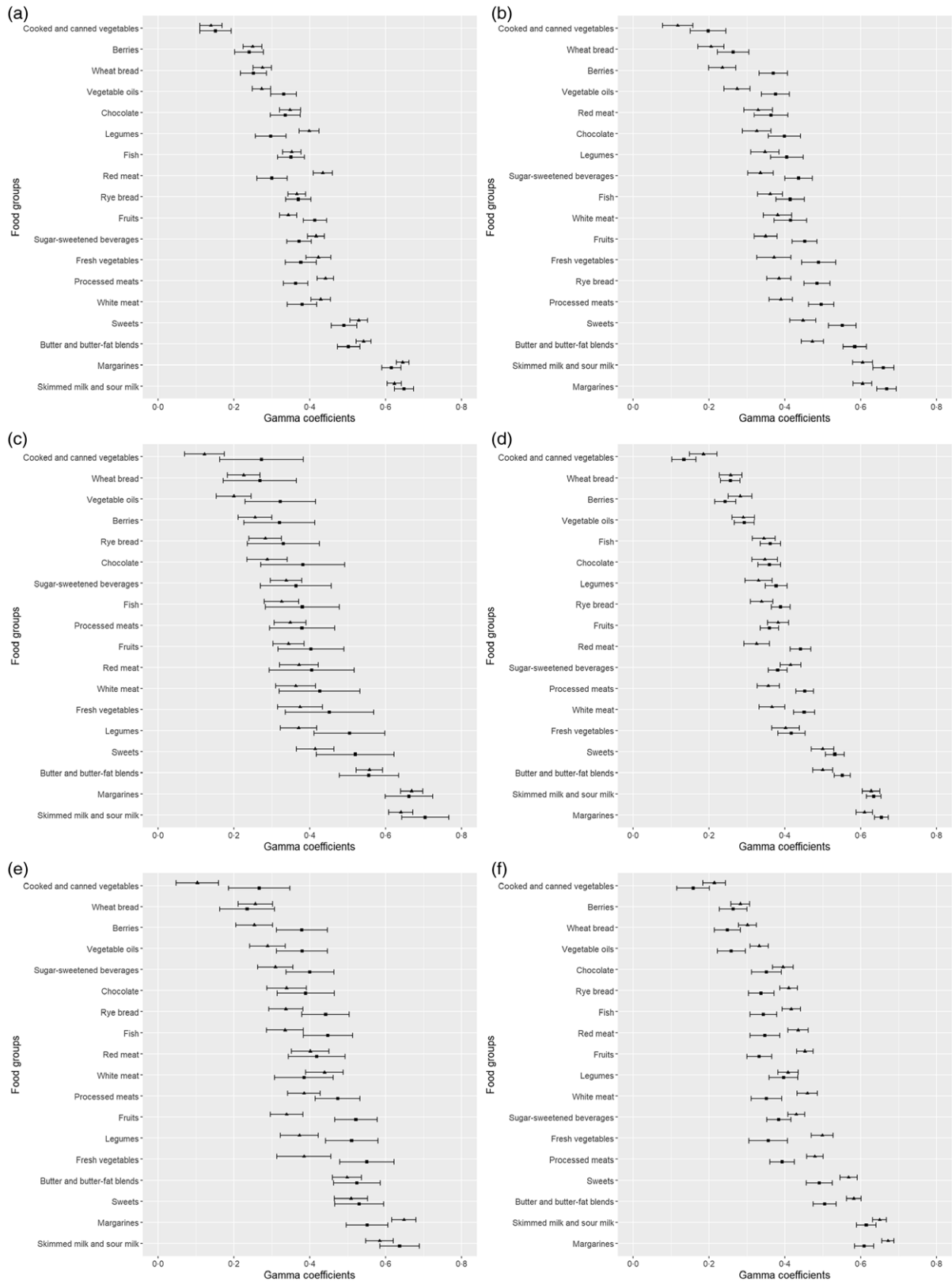
This paper describes the relative validity of loyalty-card holders' grocery purchase data as an indicator of food consumption. As the data accumulate automatically, the research process is minimally burdensome for the participants and enables recruitment of large samples, long-term follow-up and timely reactions following, for example, changes in pricing or food-related crises. Our study showed that the associations between food consumption and grocery purchase data were mostly

reasonable, suggesting that loyalty-card data on grocery purchases can be used as a proxy for food consumption. To the best of our knowledge, this is the first study to compare automatically accumulating grocery purchase data and self-reported food consumption.

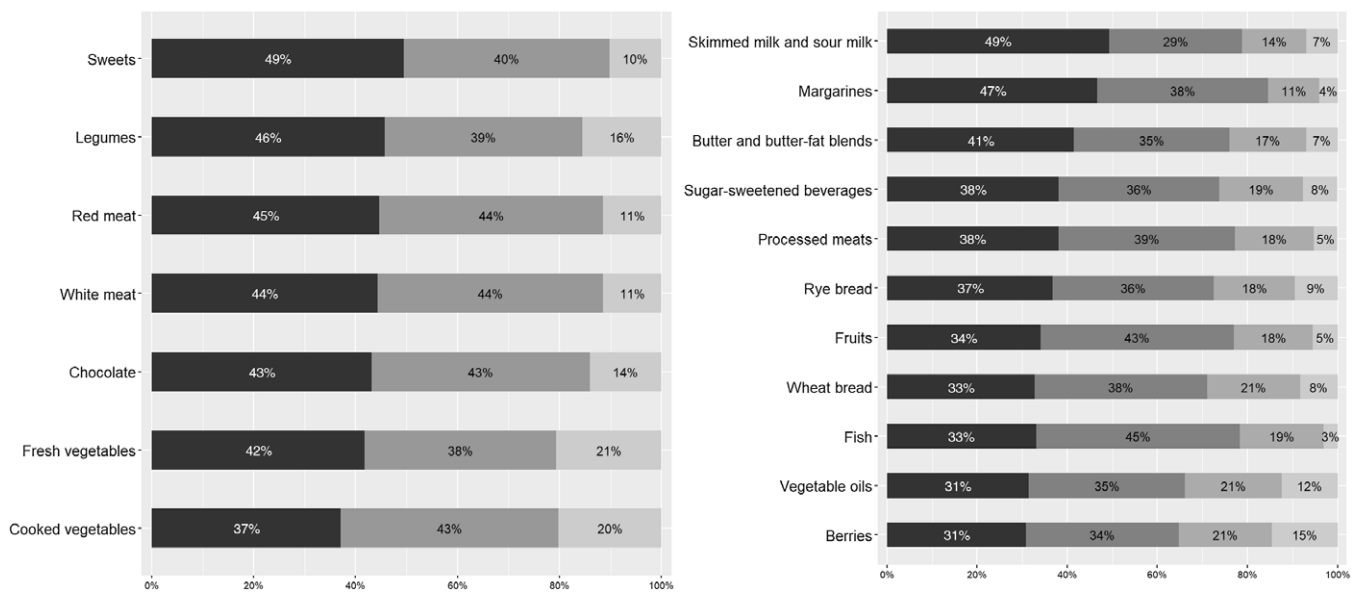
We found moderate associations (gammas ranging from 0.40 to 0.63) between grocery purchases and food consumption for skimmed milk and sour milk, margarines, butter and butter-fat blends, sweets, processed meats, fresh vegetables, white meat and sugar-sweetened beverages. The aforementioned foods are typically consumed at home as part of breakfast, dinner or evening snacks. Based on these results, it seems that the purchase volumes of these foods can, with caution, be used as a proxy for food consumption in future population-level studies, even though they cannot replace dietary assessment methods intended for extracting accurate individual-level data, such as food records or 24-h recalls. Previous studies have observed fairly similar associations using purchase records or participant-collected receipts: in a US study, the concordance between purchased and consumed vegetables was 0.39 (Lin's concordance correlation)<sup>(17)</sup>, whereas a Danish study reported a Spearman correlation of 0.28 between participant-reported vegetable purchases and a dietary quality score<sup>(20)</sup>. An earlier Swedish study reported acceptable differences (less than 20 %) between amounts of purchased and consumed foods for margarines, fish and fish dishes, and sweets, but not for vegetables, butter or milk<sup>(16)</sup>.

The observed associations in our study were somewhat weak (gammas < 0.40) for legumes, fruits, berries, vegetable oils, fish, red meat, rye bread, wheat bread and chocolate, and the





**Fig. 2.** The comparison of gammas between: men and women (a), single-adult households and two-adult families with no children (b), one- and two-adult families with children (c), participants with low and high education (d), participants with low and high scaled monthly income (e) and participants with low and high degree of loyalty (f). (a) Gender: ■, men; ▲, women. (b) Family structure: ■, single household; ▲, two adults, no children. (c) Family structure: ■, one adult and a child/children; ▲, two adults and a child/children. (d) Educational level: ■, Bachelor's degree or higher; ▲, upper secondary school or less. (e) Scaled monthly income: ■, 1000 Euros or less; ▲, 4000 Euros or more. (f) % of purchase from the retailer: ■, 60% or less; ▲, more than 60%.



**Fig. 3.** Frequency distribution (in %) of participants classified into same, adjacent (+/-1), subsequent (+/-2) and opposite categories of FFQ and 12-month purchase volume data, *n* 11 543–11 926 (food groups divided into thirds on the left, food groups divided into quarters on the right). ■, in the same third; ■, in the adjacent third (±1); ■, in the opposite third (±2); ■, in the same quarter; ■, in the adjacent quarter (±1); ■, in the subsequent quarter (±2); ■, in the opposite quarter (±3).

weakest (gamma 0.15) for cooked and canned vegetables. These results are not surprising, as for instance cooked and canned vegetables, legumes, berries, fish and red meat can be used as part of dishes, making it challenging to estimate consumption frequencies. Furthermore, larger households can have two or more food purchasers, whereas our data only covered the purchases of the main loyalty-card holder. Considering these issues, the obtained associations between purchased and consumed foods are reasonable and acceptable.

The comparison of concordances among different sub-populations yielded expected results. First, participants reporting a higher share of purchases from the retailer providing the purchase data had higher gammas. Similarly, the concordance was stronger for single-adult households than for two-adult families. The concordance for men and women was fairly similar across the food groups. Nevertheless, we observed higher gammas among women for red and processed meat, but concordance was lower for fruits. These findings could be partly related to social desirability bias: women might be more prone to responding in a manner consistent with societal norms<sup>(31)</sup>, that is, to over-report their fruit consumption. In terms of education or income, we did not find clear distinctions. However, the participants with the highest income had lower gammas for certain foods, which could be explained by the fact that they might eat out more than the participants with the lowest income. A previous study from the UK also observed a higher concordance between purchased and consumed foods among low-income households and suggested that the finding was indicative of more careful buying to meet the family's needs and less food waste<sup>(15)</sup>.

Most of the participants were classified in the same or adjacent category using the purchase volume and FFQ data. However, it is noteworthy that for foods generally considered healthy, such as fresh vegetables, cooked and canned

vegetables, berries and vegetable oils, the two methods seemed to have distinct discrepancies. Difficulties in reporting could potentially explain the observed discrepancies: for example, vegetable consumption is particularly challenging to assess using an FFQ<sup>(32)</sup>. In addition, social desirability<sup>(31)</sup>, systematic misreporting and person-specific biases related to, for example, obesity may affect the FFQ data<sup>(4,33)</sup> and thus, purchase data may reflect the consumption of these foods more realistically. Moreover, it is customary in Finland to pick berries in the forest, and thus, the mismatch between the FFQ and purchase data may plausibly reflect reality.

The purchase data used in the current study were obtained from one retailer, and we only had data from one loyalty-card holder from each household. Thus, the data used in the study did not cover all the foods bought for the household. In Finland, two large chains dominate the food retail market and in 2018 had a combined share of 83%<sup>(22)</sup>. These two chains have slightly different profiles, and people might have tendencies to buy certain foods from the other chain, resulting in possible systematic bias in our purchase data. In addition, the data only covered purchases associated with the loyalty card. Hence, it is possible that not all purchases were recorded in the data. However, it is unlikely that the participants systematically neglected to use their loyalty cards when purchasing certain foods.

The current study is one of the first to use automatically accumulating loyalty-card data on grocery purchases and could serve as a starting point for future studies investigating food behaviour in large samples. The advantages of the method include the automatic nature and objectivity of the data collection: the participants are not expected to record their food purchases, and moreover, it is difficult to consciously or unconsciously misreport purchases. Previously, loyalty-card data have been used in studies investigating, for example, customer behaviour<sup>(34–36)</sup>

and intervention effects<sup>(37–42)</sup>. An Australian study used loyalty-card data on food purchases to build an index describing the healthiness of food purchases<sup>(43)</sup>, whereas a French study used similar data to study the associations between the healthiness of food purchases and alcoholic beverages<sup>(44)</sup>. Our group also used purchase data to show that unhealthy purchases tend to cluster: the foods most consistently related to alcohol purchases were sausages, soft drinks and snacks<sup>(45)</sup>. A recent UK study used loyalty-card data to study the relationship between food purchases and hypertension, high blood cholesterol and diabetes in an ecological design<sup>(11)</sup>. Our study adds credibility to earlier studies by suggesting that at least a moderate association exists between grocery purchases and food consumption. Moreover, purchase data could be used to assess the potential effects of, for example, food policies and taxation in different age groups or areas with varying socio-demographic characteristics or educational levels, for which purposes individual-level data are not essential. Considering the imperfect match between household- and individual-level data, the relative validity criterion used in the current paper might even be regarded as unnecessarily strict.

The present study has several strengths. First, the sample is exceptionally large for a dietary assessment validation study, and we used different types of grocery purchase data (frequency and volume, 1- and 12-month periods). Similar findings irrespective of the type of purchase data suggest internal consistency. Furthermore, the findings concerning different subgroups were in accordance with our hypothesis. Second, the use of purchase data from the past 12 months evened out seasonal variation in the data. Obviously, the study also had notable limitations that need to be acknowledged. First, the study population was more educated than the general Finnish adult population: the percentage of participants with Master's degree or higher education was 24 %, whereas the corresponding percentage in the population is 11 %<sup>(23)</sup>. Second, we assessed food consumption with a relatively short FFQ that has not been validated. However, the food list in the FFQ was loosely based on a previous FFQ, which has been tested for relative validity (against food records)<sup>(24)</sup> and reproducibility (test–retest reliability)<sup>(46)</sup> among preschoolers. Furthermore, due to heavy aggregation in the FFQ and purchase data, we were forced to use thirds instead of quarters for some of the food groups, which could potentially mask important information.

### Conclusions

The current study demonstrated that for most food groups, loyalty-card data on grocery purchases are a moderately valid marker of food consumption. The associations between grocery purchases and self-reported food consumption were strongest in single-adult households and among participants with a higher degree of loyalty. Future studies should consider purchase data a resource-saving, moderately valid and objective proxy for food consumption in large samples.

### Acknowledgements

We thank the S Group for its collaboration. We are also grateful to the loyalty-card holders who provided consent for the use of their loyalty-card data in this research project.

This work was funded by The Finnish Food Research Foundation; The Finnish Foundation for Alcohol Studies; and EIT Food ('Towards a smarter shopping list' #20 041), the innovation community on Food of the European Institute of Innovation and Technology (EIT), a body of the EU, under the Horizon 2020, the EU Framework Programme for Research and Innovation.

The authors' responsibilities were as follows: H. V., J. N., S. K., J. M., S. M., L. U., M. F. and M. E. designed the study; J. N., L. U., M. F. and M. E. conducted the research; S. T. I. carried out the literature search; H. V. and J. N. analysed the data and performed the statistical analysis; H. V. wrote the paper; J. N. and M. E. had primary responsibility for the final content. All the authors reviewed and approved the final version of the manuscript.

H. V. has received a fee from the S Group. The collaboration included offering professional advice to influencers and writing a blog post with regard to interpretation of nutrition calculator in S Group's mobile app. M. F. is a member of the S Group's Advisory Board for Societal responsibility. The membership is without any compensation. The authors declare no other relationships or activities that could appear to have influenced the present work. The authors declare that they have no competing interests.

### Supplementary material

For supplementary materials referred to in this article, please visit <https://doi.org/10.1017/S0007114521004177>

### References

1. Beaglehole R, Ebrahim S, Reddy S, *et al.* (2007) On behalf of the chronic disease action group. Prevention of chronic diseases: a call to action. *Lancet* **370**, 2152–2157.
2. Brownson RC, Gurney JG & Land GH (1999) Evidence-based decision making in public health. *J Public Health Manag Pract* **5**, 86–97.
3. Willett W & Lenart E (2013) Reproducibility and validity of food frequency questionnaires. In *Nutritional Epidemiology*, 3rd ed. pp. 96–141 [W Willett, editor]. New York: Oxford University Press.
4. Slimani N, Freisling H, Illner A, *et al.* (2015) Methods to determine dietary intake. In *Nutrition Research Methodologies*, 1st ed. pp. 48–70 [JA Lovegrove, L Hodson, S Sharma, *et al.*, editors]. Chichester: John Wiley & Sons, Ltd.
5. Bennett DA, Landry D, Little J, *et al.* (2017) Systematic review of statistical approaches to quantify, or correct for, measurement error in a continuous exposure in nutritional epidemiology. *BMC Med Res Methodol* **17**, 146.
6. Westerterp KR & Goris AHC (2002) Validity of the assessment of dietary intake: problems of misreporting. *Curr Opin Clin Nutr Metab Care* **5**, 489–493.
7. Tolonen H, Helakorpi S, Talala K, *et al.* (2006) 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *Eur J Epidemiol* **21**, 409–415.
8. Centers for Disease Control and Prevention (2020) NHANES Response Rates and Population Totals Version Current. <https://www.cdc.gov/nchs/nhanes/ResponseRates.aspx> (accessed November 2020).



9. Nummela O, Sulander T, Helakorpi S, *et al.* (2011) Register-based data indicated nonparticipation bias in a health study among aging people. *J Clin Epidemiol* **64**, 1418–1425.
10. French SA, Shimotsu ST, Wall M, *et al.* (2008) Capturing the spectrum of household food and beverage purchasing behavior: a review. *J Am Diet Assoc* **108**, 2051–2058.
11. Aiello L, Schifanella R, Quercia D, *et al.* (2019) Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Sci* **8**, 1–22.
12. Nevalainen J, Erkkola M, Saarijärvi H, *et al.* (2018) Large-scale loyalty card data in health research. *Digital Health* **4**, 2055207618816898.
13. Eyles H, Jiang Y & Ni Mhurchu C (2010) Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls. *J Am Diet Assoc* **110**, 106–110.
14. Nelson M, Dyson PA & Paul AA (1985) Family food purchases and home food consumption: comparison of nutrient contents. *Br J Nutr* **54**, 373–387.
15. Ransley JK, Donnelly JK, Khara TN, *et al.* (2001) The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutr* **4**, 1279–1286.
16. Becker W (2001) Comparability of household and individual food consumption data – evidence from Sweden. *Public Health Nutr* **4**, 1177–1182.
17. Appelhans BM, French SA, Tangney CC, *et al.* (2017) To what extent do food purchases reflect shoppers diet quality and nutrient intake? *Int J Behav Nutr Phys Act* **14**, 46.
18. da Costa Louzada ML, Levy RB, Martins AP, *et al.* (2017) Validating the usage of household food acquisition surveys to assess the consumption of ultra-processed foods: evidence from Brazil. *Food Policy* **72**, 112–120.
19. de Oliveira DC, de Moura Souza A, Levy RB, *et al.* (2019) Comparison between household food purchase and individual food consumption in Brazil. *Public Health Nutr* **22**, 841–847.
20. Lund T, Watson D, Smed S, *et al.* (2017) The diet-related GHG Index: construction and validation of a brief questionnaire-based index. *Clim Change* **140**, 503–517.
21. Lintonen T, Uusitalo L, Erkkola M, *et al.* (2020) Grocery purchase data in the study of alcohol use – a validity study. *Drug Alcohol Depend* **214**, 108145.
22. Finnish grocery trade association (2019) Päivittäistavara-kaupan myynti ja markkinaosuudet 2018 (Sales and Market Share of the Grocery Trade Groups in 2018). [https://www.pty.fi/fileadmin/user\\_upload/tiedostot/Julkaisut/Vuosijulkaisut/EN\\_2019\\_vuosijulkaisu\\_lr.pdf](https://www.pty.fi/fileadmin/user_upload/tiedostot/Julkaisut/Vuosijulkaisut/EN_2019_vuosijulkaisu_lr.pdf) (accessed November 2020).
23. Vuorinen A, Erkkola M, Fogelholm M, *et al.* (2020) Characterization and correction of bias due to nonparticipation and the degree of loyalty in large-scale Finnish loyalty card data on grocery purchases: cohort study. *J Med Internet Res* **22**, e18059.
24. Korkalo L, Vepsäläinen H, Ray C, *et al.* (2019) Parents' reports of preschoolers' diets: relative validity of a food frequency questionnaire and dietary patterns. *Nutrients* **11**, 159.
25. Official Statistics of Finland (OSF) (2020) Dwellings and Housing Conditions (e-Publication). Appendix Table 1. Household-Dwelling Units by Number of Person 1960–2019. [https://tilastokeskus.fi/til/asas/2019/asas\\_2019\\_2020-05-20\\_tau\\_001\\_en.html](https://tilastokeskus.fi/til/asas/2019/asas_2019_2020-05-20_tau_001_en.html) (accessed September 2021).
26. Official Statistics of Finland (OSF) (2021) Families (e-publication). Appendix Table 1. Families by Type in 1990–2020. [https://www.stat.fi/til/perh/2020/perh\\_2020\\_2021-05-28\\_tau\\_001\\_en.html](https://www.stat.fi/til/perh/2020/perh_2020_2021-05-28_tau_001_en.html) (accessed September 2021).
27. OECD (2020) Compare your income – Methodology and Conceptual Issues. <https://www.oecd.org/statistics/Compare-your-income-methodology.pdf> (accessed November 2020).
28. Agresti A (2013) *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: Wiley.
29. BMJ Correlation and Regression. (2020) <https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression> (accessed March 2021).
30. Revicki D (2014) Internal consistency reliability. In *Encyclopedia of Quality of Life and Well-Being Research*, pp. 3305–3306 [AC Michalos, editor]. Dordrecht, The Netherlands: Springer Netherlands.
31. Hebert JR, Ma Y, Clemow L, *et al.* (1997) Gender differences in social desirability and social approval bias in dietary self-report. *Am J Epidemiol* **146**, 1046–1055.
32. Cade JE, Burley VJ, Warm DL, *et al.* (2004) Food-frequency questionnaires: a review of their design, validation and utilisation. *Nutr Res Rev* **17**, 5–22.
33. Lissner L (2002) Measuring food intake in studies of obesity. *Public Health Nutr* **5**, 889–892.
34. Sturley C, Newing A & Heppenstall A (2018) Evaluating the potential of agent-based modelling to capture consumer grocery retail store choice behaviours. *Int Rev Retail Distrib Consum Res* **28**, 27–46.
35. Paulssen M, Bruneder J & Sommerfeld A (2019) Customer in-role and extra-role behaviours in a retail setting. *Eur J Mark* **53**, 2501–2529.
36. Lloyd A & Cheshire J (2019) Detecting address uncertainty in loyalty card data. *Appl Spat Anal* **12**, 445–465.
37. Sharma A, di Falco S & Fraser I (2019) Consumption of salt rich products: impact of the UK reduced salt campaign. *Int J Health Econ Manag* **19**, 341–357.
38. Polacsek M, Moran A, Thorndike AN, *et al.* (2018) A supermarket double-dollar incentive program increases purchases of fresh fruits and vegetables among low-income families with children: the Healthy Double Study. *J Nutr Educ Behav* **50**, 217–228.e1.
39. Piernas C, Tsiountsioura M, Astbury NM, *et al.* (2019) Primary care shopping intervention for cardiovascular disease prevention (PC-SHOP): protocol for a randomised controlled trial to reduce saturated fat intake. *BMJ Open* **9**, e027035.
40. Moran AJ, Khandpur N, Polacsek M, *et al.* (2019) Make it fresh, for less! A supermarket meal bundling and electronic reminder intervention to promote healthy purchases among families with children. *J Nutr Educ Behav* **51**, 400–408.
41. Harrington RA, Scarborough P, Hodgkins C, *et al.* (2019) A pilot randomized controlled trial of a digital intervention aimed at improving food purchasing behavior: the front-of-pack food labels impact on consumer choice study. *JMIR Form Res* **3**, e9910.
42. Franckle RL, Levy DE, Macias-Navarro L, *et al.* (2018) Traffic-light labels and financial incentives to reduce sugar-sweetened beverage purchases by low-income Latino families: a randomized controlled trial. *Public Health Nutr* **21**, 1426–1434.
43. Taylor A, Wilson F, Hendrie GA, *et al.* (2015) Feasibility of a healthy trolley index to assess dietary quality of the household food supply. *Br J Nutr* **114**, 2129–2137.
44. Hansel B, Roussel R, Diguët V, *et al.* (2015) Relationships between consumption of alcoholic beverages and healthy foods: the French supermarket cohort of 196 000 subjects. *Eur J Prev Cardiol* **22**, 215–222.
45. Uusitalo L, Erkkola M, Lintonen T, *et al.* (2019) Alcohol expenditure in grocery stores and their associations with tobacco and food expenditures. *BMC Public Health* **19**, 787.
46. Määttä S, Vepsäläinen H, Lehto R, *et al.* (2018) Reproducibility of preschool personnel and guardian reports on energy balance-related behaviors and their correlates in Finnish preschool children. *Children* **5**, 144.