Measuring domestic food safety:

A review of the Index of Recommended Practice

Social Science Research Unit

Food Standards Agency

FS409012

August 2016

Unit Report 2016/01

NatCen Social Research

NatCen
Social Research that works for society

Measuring domestic food safety:

A review of the Index of Recommended Practice

Authors: David Hussey[1] Caireen Roberts[1] Laura Inman[3] Michael Howard[2] Sally McManus[1]
[1]NatCen Social Research
[2] Kings College London
[3] Food Standards Agency
Date: August 2016
Prepared for: Food Standards Agency

At **NatCen Social Research** we believe that social research has the power to make life better. By really understanding the complexity of people's lives and what they think about the issues that affect them, we give the public a powerful and influential role in shaping decisions and services that can make a difference to everyone. And as an independent, not for profit organisation we're able to put all our time and energy into delivering social research that works for society.

NatCen Social Research
35 Northampton Square
London EC1V 0AX
T 020 7250 1866
www.natcen.ac.uk

A Company Limited by Guarantee
Registered in England No.4392418.
A Charity registered in England and Wales (1091768) and Scotland (SC038454)

Contents

Executive Summary

Food and You is a biennial, random probability, cross-sectional survey of approximately 3,000 adults (16 years and over) living in private households in England, Wales and Northern Ireland, commissioned by the Food Standards Agency (FSA). The survey includes questions on reported behaviour, knowledge and attitudes relating to food safety issues. So far, three waves of Food and You have been completed (in 2010, 2012 and 2014). In Wave 2, a composite measure of reported domestic food safety practices, the Index of Recommended Practice (IRP), was developed as a tool to get more information about which socio-economic and demographic groups were least likely to report behaviour in line with recommended practice. After the publication of the Wave 2 report in 2013, the FSA redeveloped the IRP so that it could track progress between Wave 1 and Wave 2. This re-developed version comprised 17 questions, and respondents were allocated a score of 1 for responses in line with FSA Recommended Practice (RP) or 0 for responses not in line with Recommended Practice (non-RP).

A peer review of the re-developed IRP was commissioned in 2013 by the FSA, using data from Waves 1-2 of the Food and You, in order to specifically evaluate it against the stated aims of tracking progress towards improving public awareness and use of messages about good food hygiene practice at home, and gathering data on domestic food safety practices in order to inform future policies on food safety.

Following a qualitative and quantitative peer review, further amendments to the IRP were considered. This report focuses on deliberation of the changes to the IRP following suggestions from the reviews, rationale for changes made, and details of the implementation of changes and composition of the revised IRP.

A deliberative workshop was held to discuss potential changes to the IRP. Changes to the IRP were recommended in response to the reviews and recommendations from the workshop. The main changes were:

- Questions that measure knowledge should either be removed from the new IRP entirely or combined with the related behaviour, in order to address the issue of ascribing high scores to respondents who self-reported knowledge of Recommended Practice but who did not report the associated behaviour
- Groups or pairs of questions that measure the same form of behaviour should be combined to reduce the duplication of information
- A threshold should be applied for the number of questions a respondent must answer before a score can be derived, in order to facilitate comparisons between respondents

The new (revised) IRP is now constructed using 16 questions from Food and You. These questions can be used to measure 10 domestic food safety behaviours. Each item is scored 1 for RP responses or 0 for non-RP. Respondents answering less than half (five) of the 10 items do not receive a score.

With the changes made to the construction of the index following the review, we can have confidence that the IRP is a robust tool which can be used to explore domestic food safety practices and track changes in these practices over time.

1 Introduction

1.1 Background

Improving awareness and use of messages about good food hygiene behaviour in the home was a key aim in the Food Standards Agency's (FSA) 2010-2015 strategy. In order to monitor progress towards this aim, and to provide general insight into domestic food safety practices, the FSA commissioned the Food and You survey. Food and You is a biennial, random probability survey, currently conducted in England, Wales and Northern Ireland with approximately 3,000 respondents at each wave.[1] Three waves of Food and You have been completed to date (in 2010, 2012 and 2014), with a fourth wave (2016) currently in progress. At Wave 2, a composite measure of reported domestic food safety practices, the Index of Recommended Practice (IRP), was developed as a tool to get more information about trends among socio-economic and demographic groups reporting behaviour in line with recommended practice. This original IRP included 14 questions taken from the Wave 2 questionnaire covering five domains of food safety practice: chilling, cooking, cleaning, cross-contamination and use by dates. The questions were selected because they mapped onto practices that, if not followed, were likely to increase the risk of exposure to food-borne disease causing bacteria. After the publication of the Wave 2 report in 2013, the FSA redeveloped the IRP as a tool that could be used to:

- Track progress towards its aim of 'improving public awareness and use of messages about good food hygiene practice at home' over the course of their 2010 – 2015 strategy and beyond.
- Increase the FSA's understanding of domestic food safety practices in order to inform policy and communication strategies.

Redevelopment of the IRP into its current form involved changing some questions so that the measure could track progress across the Wave 1 and Wave 2 Food and You data, as some of the questions in the original composite were only included in the Wave 2 survey. Respondents were allocated a score for reporting practices that were in line with FSA recommendations for the following 17 questions:

- Do you ever check your fridge temperature?
- How often do you or another person in your household check the temperature of the fridge?
- Thinking about fridge temperature, can you tell me how you normally check the temperature?
- What do you think the temperature inside your fridge should be?
- Do you do the following things at all when you are in the kitchen and if so how frequently: cook food to steaming hot?
- Do you do the following things at all when you are in the kitchen and if so how frequently: eat chicken or turkey if the meat is pink or has pink or red juices?
- How many times would you consider re-heating food after it was cooked for the first time?

[1] At Waves 1-3, the Food and You survey was conducted with a UK-wide sample. In April 2015 the FSA's responsibilities in Scotland were transferred to the new independent Scottish food safety body, Food Standards Scotland (FSS). As a result, Wave 4 of Food and You is being carried out in England, Wales and Northern Ireland only. This review was conducted using data from Waves 1-2 of the survey, and findings therefore relate to aggregate UK-level data from all four countries.

- And how do you usually tell that food has been reheated properly?
- Do you do the following things at all when you are in the kitchen and if so how frequently: wash raw meat and poultry?
- Why do you think people wash chopping boards after preparing raw meat, poultry or fish?
- Where in the fridge do you store raw meat and poultry?
- How do you store raw meat and poultry in the fridge?
- Do you do the following things at all when you are in the kitchen and if so how frequently: wash hands before starting to prepare or cook food?
- Do you do the following things at all when you are in the kitchen and if so how frequently: wash hands after handling raw meat / fish?
- Which of these [date labels] indicates whether food is safe to eat?
- Do you check use by dates when you are about to cook or prepare food?
- If you made a meal on Sunday, what is the last day that you would consider eating the leftovers?

Details of the response codes for the above questions and derivation of the IRP can be found in Appendix A.

1.2 Aims of this report

A peer review of the current IRP was commissioned in 2013 by the FSA to specifically evaluate it against the stated aims to track progress and gather data to inform understanding. The peer reviewers were asked to evaluate and comment from both a qualitative and a quantitative perspective. Their recommendations were then discussed at a workshop and further refinements were made to the IRP.

This report focuses on the following stages of the project:

1. Deliberative workshop and recommendations
2. Resulting changes to the scoring for a revised IRP

An overview of the initial peer review components which informed the workshop can be found in Appendix B.

2 Suggestions following peer review

The IRP was reviewed both quantitatively and qualitatively (see Appendix B for an overview of the peer review comments). This review broadly found that the current IRP performs well in terms of the evaluation criteria. There were four main areas for further consideration (with some overlap):

- Knowledge vs. behaviour
- Weighting of the data
- Missing data
- Binary vs. ordinal responses

An important consideration when discussing these issues was that any construction of the IRP is straightforward to explain and use. In introducing possible refinements, we introduce added complexity which makes the index harder to understand and more likely to be used or interpreted incorrectly by other users.

2.1 Knowledge vs. behaviour

The current composite mainly contains questions on reported behaviour, with a few knowledge questions. Reported behaviour is important as it could provide insight into the domestic food safety practices that respondents undertake. However, knowledge could underpin the behaviours. This raises the question of whether responses on knowledge should be treated the same as reported behaviour, and also whether knowledge in isolation is a desirable measure, or whether knowledge is only valuable if it underpins a specific reported behaviour. If the latter then we suggest knowledge questions only receive a score when associated behaviour is in line with recommended practice. Otherwise, we would also suggest considering re-labelling the index so that it is clear that 'knowledge' is covered as well as reported behaviour.

2.2 Weighting the data

There are two sets of weights which could be applied to the index to be considered: question-level weighting, and domain-level weighting. Currently, the unequal number of questions within each domain means that there is effectively a domain-level weight, as greater weight is put on the domains which contain more questions. Whether this is desirable should be considered, or instead should all domains receiving equal weighting?

There is currently no question-level weighting but some of the questions address issues that could be considered more important to food safety than others. This raises the issue of whether questions should be individually weighted according to food safety 'importance'.

Missing data would have implications for both the domain-level and question-level weights, as questions which represent missing data are currently excluded from the respondent's base. This issue is detailed in the section below.

2.3 Missing data

Missing data currently occur when a question is not applicable to a respondent. If many questions are not applicable to a respondent, this could suggest that they rarely undertake

the domestic food safety practices that the index considers. It is then arguably not useful to allocate a score to these participants using the index. With regard to the scoring it is important to consider three aspects: (i) the number/proportion of questions answered overall; (ii) the number answered within each domain; and (iii) the pattern of response across domains. Whether respondents should be excluded if they do not meet a threshold number of questions on any or all of these aspects should be considered.

The current overall score is calculated to take into account questions that are not applicable / missing by giving respondents a base of valid questions. If question-level weights are added, respondents could effectively 'lose out' on gaining a higher score if the questions that carry more weight are not applicable to them. Alternatively, the value of questions with lower weights will be falsely inflated for these respondents. The effect of weighting on missing data, and how this could be dealt with, should also be considered.

2.4 Binary vs. ordinal responses

Finally we consider the relative strengths and weaknesses of using binary responses (recommended practice vs non-recommended practice) compared with ordinal (for example a rating scale from 'always follows RP' to 'never follows RP').

The current IRP allocates a point (+1) to reported practices which are in line with Agency recommendation, and zero to reported practices which are not in line. However this means that all practices which fall outside of the FSA's guidance are treated the same i.e. awarded a zero. The qualitative review flagged a few examples such as, while knowing that the fridge should be below 5°C is in line with recommended practice, saying 5-8°C is closer to being in-line with recommended practice, and also likely to be less 'risky' in terms of food safety than thinking it should be higher than 8°C. The qualitative review also suggested that checking the fridge temperature four times per year "would produce a huge improvement to food safety compared to 'never'", although both of these responses are not in-line with Agency guidance.

Including more detail like this may make the composite measure more sensitive to identifying change. For example, in 2010 a respondent may think the fridge should be at 10°C and in 2014 they think it should be at 8°C. Currently their improvement is not captured as both are not in line with Agency advice. The pros and cons of adopting ordinal responses are shown in Table 2.1.

Table 2.1

Pros	Cons
• More sensitive to changes in knowledge and behaviour over time • More sensitive to picking up on differences between groups • Less loss of detail	• Makes scoring more complicated and the index harder to explain and use • It would no longer be a measure of pure adherence to 'recommended practice'

3 Recommendations from deliberative workshop

The four main areas for consideration were taken forward to a deliberative workshop. Attendees were made up of the NatCen project team, representatives from the FSA's Social Science Research Unit, Operational Research Unit, Statistics Branch, Hygiene and Microbiology Division, Consumer Engagement, and expert members of the Food and You Working Group, external to the FSA. To focus the discussion, attendees were asked a series of questions under the umbrella of each of the four areas (shown in the boxes below). The questions and notes of the subsequent discussions are reported in the following sections.

3.1 Knowledge vs. behaviour

The current IRP mixes questions on knowledge and behaviour. The aim of the IRP is to track progress towards the FSA's strategic priority to improve public awareness and use of messages about good food hygiene practice at home.

Questions:
- Given that the role of the IRP is to measure food safety 'practice', what should the role of knowledge questions be?
- Should knowledge questions be included at all?
- Should responses on knowledge be scored only when linked with recommended behaviour?

The issue of whether the IRP should include knowledge and behaviour questions is related to the definition of the IRP. There is currently a definition ('The index measures the extent to which reported food safety behaviour was in line with Agency recommended practice') and an aim of the IRP (to 'track progress towards the Agency's strategic priority to 'improve public awareness and use of messages about good food hygiene practice at home''). These do not appear to be harmonious – the definition suggests that the IRP should measure behaviour and whether this is in line with recommended practice. The aim suggests that both behaviour and knowledge are important in the composite. It will be important in future to make sure that the definition and aim of any form of the IRP is clear and correct for purpose.

There are three knowledge questions in the composite. One ('What do you think the temperature inside your fridge should be?') is necessary in order to underpin whether the reported behaviour is in line with recommended practice (If a respondent checks the temperature of their fridge, they generally need to know what the temperature should be in order to do this effectively). The other two knowledge questions (one on cross-contamination and one on use by dates) do not underpin behavioural questions in the same way,[2] and could be argued to be outside the definition of the IRP (as knowledge does not necessarily equate to or underpin recommended practice). One option is that these two questions could be dropped from the current composite to create a composite which measures reported behaviour. Workshop attendees considered this to be a reasonable option as it was thought

[2] After further consideration of this issue following the workshop, we decided the use by dates question does underpin behaviour to some extent

that overall behaviours are more important for the composite to measure than knowledge as it is the behaviour which ultimately leads to reducing risk.

There are however arguments for retaining questions on knowledge in the IRP. Questions on knowledge can be complementary to those on behaviour and may be useful in underpinning behaviour. Workshop attendees suggested that one option to take account of knowledge would be to develop separate 'knowledge' and 'behaviour' composites. The knowledge composite would contain an array of the knowledge questions from Food and You (but would be dependent on the number of suitable questions and whether these were present in both Wave 1 and Wave 2). The behaviour composite would include those questions currently in the composite with the two stand-alone knowledge questions removed. The chilling questions are a special case and it could still be necessary to include the fridge temperature question to underpin the behaviour.

As there are only 3/17 questions in the index which relate to knowledge, it was thought that if knowledge questions were considered to be an important part of the IRP, a case could be argued for including more knowledge questions.

RECOMMENDATION: Review the knowledge questions with a view to the IRP being built on 'behaviours'. Consider chilling knowledge question separately as this underpins behaviour.

3.2 Weighting the data

In the current IRP, individual questions are not weighted i.e. each question is treated as equally important/valid. As the domains do not have an equal number of questions there is an 'unintentional' weighting of the domains.

Questions:
- What are the pros and cons of weighting or not weighting the individual questions?
- If questions were individually weighted, what weights should be given to each?
- Should the five domains carry equal weighting?
- How important is it to retain simplicity in the scoring design?

The questions in the current IRP are un-weighted. This means that the five domains (4 Cs plus use by dates) are weighted differently as there are different numbers of questions relating to each domain (2 cleaning, 4 cooking, 4 chilling, 4 cross contamination, 3 use by dates). The issue of whether the questions or domains should be weighted, and how, is complex, particularly as the questions and domain structures are not independent. It was thought that if the domains are important then they should each have the same weight (the only reason one is currently weighted more than another is because we happen to have more suitable questions on that domain). But if the domains were weighted equally (to account for the unequal question numbers), this will effectively weight the questions differently (i.e. answering a cleaning question will be 'worth more' than answering a cross contamination question). Therefore, it is necessary to consider which is more important in the construction of the composite: the questions or the domains. Weighting one will necessarily affect the other.

The importance of the domains was discussed, and there was split opinion on their usefulness. The 5 domains are useful in terms of delivering messages. However for analytical purposes the domains could be considered artificial, without inherent values, in

which case it would be sensible to ignore the domains in the construction of the IRP. Workshop attendees thought that it is perhaps more important for the questions to be weighted equally than the domains. If, in the future, there is interest in one particular domain, or it becomes important to consider the output in terms of domains, then each domain could be considered independently from the main IRP as necessary.

As well as discussing the consequences of inadvertent weighting due to the unbalanced number of questions, whether questions / domains should be purposefully weighted in terms of food safety 'importance' was also considered.

The five domains were not thought to be equally important, but opinion varied over the order of importance of the domains. It was thought that individual questions could be weighted in terms of 'food safety' using scientific principles, i.e. which behaviours are most likely to lead to risk. However, it was considered that this may not be worthwhile if all questions were rated reasonably highly. In addition to this, although weighting in terms of food safety 'importance' would lead to greater subtlety, it was thought that what this would ultimately achieve should be considered prior to making the IRP more complex by introducing weighting. Any decisions around weighting should be clearly documented so that the IRP is transparent and robust, including consideration of the consequences of weighting and how this may change the meaning of the IRP. A key argument against weighting was that the IRP should be kept as simple as possible, particularly for communication and interpretation, and also considering full analysis of key food safety issues is likely with each additional release of data.

RECOMMENDATION: Consideration to be given to weighting individual questions and removing the domains in the construction of the IRP.

3.3 Missing data

In the current IRP, overall score = total number of questions answered as RP/ total number of questions answered.

Questions:
- Should there be a minimum number of questions answered overall in order for a respondent to receive a score?
- Should there be a minimum number/proportion of questions answered within a domain in order for a respondent to receive a score?
- Should respondents have to answer questions in all five domains in order to receive a score?
- Should 'Don't Know' always be scored as '0 = Not RP'?

All responses are currently coded. If a question is 'not applicable' to a respondent, then that question is removed from their base. This raised the question of whether there should be a minimum number of questions that respondents should answer in order to receive a score.

In general it was thought that a respondent should answer a minimum number of questions to receive a score. This is reasonable as the only way that respondents have missing data is if the questions are not applicable to them and it is perhaps not relevant to consider the domestic food safety habits of people who e.g. do not regularly prepare food. What the minimum number of questions answered should be is a subjective issue. The best way to

decide this may be to look at the data to find a suitable cut-off point which balances the number of questions answered and number of respondents excluded.

It was also considered whether there should be a minimum number of questions answered within a domain, and whether questions needed to be answered in each domain in order for a respondent to receive a score. This is further complicated by the unequal numbers of questions across domains, so there is no obvious solution. If it is decided that the domains are a key part of the IRP, then it was thought that questions in all 5 domains should be answered in order to receive a score, although the minimum number of responses in each domain would need to be considered further.

Participants are able to answer 'don't know' (DK) to each question, and this is currently scored as 0 = Not RP. A DK response could indicate that the participant does not want to answer the question, or that they genuinely don't know or are unable to recall their practices.

Whether 'don't know' responses should be scored as '0 = Not RP' was discussed. It was thought that further analysis to explore the number and profile of 'don't know' responses would be a good place to start in addressing this issue. If there are very few DK responses, then this is unlikely to be an issue that needs to be considered in detail, and excluding these responses could be considered (although in general it was felt that all responses should be scored as far as possible). If there are a large proportion of DK responses, it will be necessary to explore the profiles of these respondents to attempt to understand these responses further.

As a DK response could indicate that the respondent is unable to recall a habitual behaviour, it was felt that DK responses should be looked at on a question by question basis as some may be more important, or could be considered to have different meanings, to others. It was thought that a DK response could still represent behaviour in line with RP (as the respondent may not be able to recall a habitual behaviour). Food and You is not sensitive enough to tease this out, but it could be used as an argument for excluding DK responses in some cases (although it would be necessary to consider each question separately). It was however thought that the DK responses were currently dealt with in a parsimonious way, and that it would be necessary to justify any changes made to how they are scored.

RECOMMENDATION: Introduce a minimum number of questions that a respondent needs to answer in order to receive a score. Conduct analysis of patterns of NA and DK responses to inform the decision.

3.4 Binary vs. ordinal

In the current IRP, responses are binary (1=RP; 0=Not RP).

Questions:
- Should the IRP remain a measure of only recommended practice with binary scores for each question or use a sliding scale that also recognises 'better' practice?
- How would the questions be scored on a sliding scale and who would decide how the scores are calculated?

A key benefit of the current binary measure is that it is simple, which facilitates explaining and understanding outcomes relating to the IRP. There was clear support for keeping the IRP as simple as possible. However, changing to an ordinal measure of recommended

practice would increase sensitivity, allowing for other changes to be detected over time (e.g. never washing raw meat and poultry is RP. Sometimes washing is arguably less dangerous than always washing, but the current IRP does not account for this and would score both as non-RP).

There are issues with introducing an ordinal scale which need to be considered. Introducing an ordinal scale would introduce subjectivity into the IRP as it will be necessary to 'score' the different responses (i.e. it would be necessary to quantify how much closer to RP washing meat sometimes is compared to always). This is a complex issue and it was thought the most pragmatic solution would be to implement equal intervals between the scores.

RECOMMENDATION: Binary model to be retained but consideration to be given to an ordinal model in the future to allow greater sensitivity when tracking changes over time.

4 Revisions to the current IRP

Following the workshop, and with on-going consultation with social science, operational research and microbiology staff at the FSA, changes were made to the scoring of the IRP to address some of the issues raised by the review and discussed at the workshop.

In summary these were:

- Modification to the treatment of knowledge questions – these questions were either removed from the new IRP entirely or combined with the related behaviour. This is discussed in section 4.1
- Combining of groups/pairs of questions that measure one behaviour. This is discussed in section 4.2
- Establishing a threshold of number of questions answered below which respondents do not receive a score on the new IRP. Analysis of the patterns of 'not applicable' (NA) and 'don't know' (DK) responses is discussed in section 4.3

There was a clear case for making all the above changes and details are given in the relevant sections. These changes were also relatively simple to implement. The 5 domains have been kept for descriptive purposes only (equal domain weighting was not adopted). Two areas which required more consideration were a) weighting individual questions and b) introducing an ordinal scale for questions. The final decision was not to weight the questions according to 'risk' based on the fact that this would need a judgement on how 'important' each practice was in terms of food safety. This would be a subjective judgement and therefore difficult to justify and reach consensus on – particularly as the 'most important' would be likely to change as the FSA's priorities changed and is likely to differ between individuals. Not weighting the questions also has the advantage of keeping the IRP simpler for users and less likely to be interpreted incorrectly. Furthermore, as the changes to the IRP include combining groups / pairs of questions that measure one behaviour, it would be necessary to identify a whole behaviour (e.g. chilling) as being more 'risky' than another behaviour, rather than assessing the 'risk' of an element of behaviour addressed by a single question (e.g. checking fridge temperature). Although it may be possible to identify individual questions as more or less 'important' in terms of food safety, it is not possible to categorise the overarching behaviours in this way.

As a result of these changes, the new (revised) IRP comprises ten items, each scored 1 for responses in line with Recommended Practice (RP) or 0 for responses not in line with Recommend Practice (non-RP). Each item is derived either from individual questions on Food and You or from pairs of questions or (in one case) a group of four questions. Respondents answering less than half (five) of the ten items do not receive a score on the revised IRP.

The revised IRP is detailed in Appendix D.

4.1 Knowledge questions

The following changes to the composite were made for these questions:

Q4.3 ***Why do you think people wash chopping boards after preparing raw meat, poultry or fish?***

This question doesn't directly measure a behaviour nor does it underpin any other measured behaviour in the IRP. Therefore it was agreed that it was sensible to remove it from the IRP.

Q4.19 Which of these indicates whether food is safe to eat?
Q4.22 Do you check use by dates when you are about to cook or prepare food?

We agreed that Q4.19 and Q4.22 together measure one behaviour. Logically respondents must check use by dates and know that they indicate whether food is safe to eat in order to score any points. Therefore respondents must follow RP for both questions to get a score of 1.

Q4.12 What do you think the temperature inside your fridge should be?

See section 4.2.1

4.2 Combining groups/pairs of questions

4.2.1 Chilling questions

As discussed in the workshop, while the four chilling questions included both knowledge and behaviour, the knowledge underpins the behaviour and therefore it was agreed that these questions together measure one behaviour. Logically respondents must check the fridge temperature in the correct manner and know what it should be in order to score any points. A parsimonious approach was agreed whereby respondents must follow RP for all four questions to receive a score of 1 and zero otherwise. This does have the disadvantage that it is quite difficult to score a 1 with this new system (only 15% of respondents from the Wave 1 and 2 dataset do).

4.2.2 Storing raw meat and poultry in the fridge

Q4.14 Where in the fridge do you store raw meat and poultry?
Q4.15 How do you store raw meat and poultry in the fridge?

These both measure how food is stored in the fridge. Although the two questions measure different aspects, arguably this is only one behaviour and in order for the behaviour to be "safe" both aspects should follow RP. For example, it is not safe to store raw meat on the bottom shelf if it isn't covered and/or away from cooked foods. We therefore agreed that respondents would receive a score of 1 only if both questions followed RP.

4.2.3 Washing hands

Q4.1.11 Do you do the following things at all when you are in the kitchen and if so how frequently: Wash hands before I start preparing or cooking food.
Q4.1.12 Do you do the following things at all when you are in the kitchen and if so how frequently: Wash hands after handling raw meat/fish

These both measure frequency of hand washing. Although the two questions measure different risks associated with hand washing (Q 4.1.11 captures the risk of the environment contaminating the food, whereas Q4.1.12 captures the risk of the food contaminating the environment), this is arguably only one behaviour. Given the changes to the earlier questions, it was decided that it would make sense to score these together and for both to have to follow RP in order for a respondent to receive a score of 1. Moreover, almost 90% of

those who said they “always” wash their hands after handling raw meat/fish also said they “always” wash their hands before food preparation.

4.3 Patterns of ‘not applicable’ and ‘don’t know’ responses

In the construction of the current IRP, NA responses are scored as missing. This results in the question being excluded from the calculation of the IRP score for that individual respondent. DK responses on the other hand are scored as not RP (0). The logic here is that DK indicates the behaviour applies to the respondent (as otherwise they would answer NA) and as they cannot be scored as RP they are, by default, scored as not RP. Currently, no consideration has been given to the number or proportion of DK and/or NA responses given by individual respondents across the questions that make up the IRP.

In the analysis conducted as part of the quantitative peer review, it was noted that a small proportion of respondents gave NA responses to a majority of the 17 items (or equivalently had an “overall base” of 8 or less - see section 3.4). It was suggested that there was an argument to be made for not scoring individuals who answered below a certain threshold number of questions. Furthermore, in the workshop it was pointed out that little was known about the extent or pattern of NA (and DK) responses.

It was therefore decided that the revision of the current IRP provided an ideal opportunity to review the patterns of these answers in the data from Wave 1 and Wave 2 of Food and You, and that this might potentially inform a revised approach to scoring the IRP (there was no specific proposal to change the treatment of DK responses rather a feeling that more needed to be understood about their distribution). The analysis used the new ten-item IRP. The ten items are listed in Table 4.1.

Table 4.1

Item	Question(s)
1	Chilling questions (Q4.9-Q.12)
2	Cooking food to steaming hot (Q4.1.13)
3	Eating chicken/turkey if meat is pink or has pink/red juices (Q4.1.14)
4	Number of times you would consider re-heating food (Q4.25)
5	How you usually tell food has been re-heated properly (Q4.26)
6	Washing raw meat/poultry (Q4.1.5)
7	Where/how you store raw meat and poultry in the fridge (Q4.14/Q4.15)
8	Washing hands before food preparation/after handling raw meat/fish (Q4.1.11/Q4.1.12)
9	Knowledge and checking of use by dates (Q4.19/Q4.22)
10	Last day you would consider eating Sunday leftovers (Q4.24)

4.3.1 Patterns of 'not applicable' (NA) responses

The number of NA responses per item is shown in Table 4.2. Note that two items (9 and 10) have no respondents in this category and a third (chilling) has a fairly negligible number (25).

Table 4.2

Item	Question(s)	No of N/A responses	Percent
5	How usually tell food has been re-heated	1,191	18.6%
7	Where/how store raw meat/poultry in the fridge *	654	10.2%
6	Washing raw meat/poultry	586	9.2%
4	No of times would consider re-heating food	509	8.0%
3	Eating meat with pink/red juices	277	4.3%
2	Cooking food to steaming hot	211	3.3%
8	Washing hands **	129	2.0%
1	Chilling questions	25	0.4%
9	Knowledge and checking of use by dates	0	0.0%
10	Last day would consider eating leftovers	0	0.0%

* for this item, NA responses on *either* question result in a NA score for the item
** for this item, only NA responses on *both* questions result in a NA score for the item i.e. where one question is NA and one isn't the item is scored on the one applicable question.

The distribution of NA responses across the ten items (and the equivalent overall base) is shown in Table 4.3.

Table 4.3

Total NA responses	Overall Base	Number of respondents	Percent	Cumulative percent
7	3	5	0.1%	0.1%
6	4	10	0.2%	0.2%
5	5	38	0.6%	0.8%
4	6	80	1.3%	2.1%
3	7	261	4.1%	6.2%
2	8	586	9.2%	15.3%
1	9	1,022	16.0%	31.3%
0	10	4,392	68.7%	100.0%

Just over two-thirds of respondents answer all ten questions (i.e. do not give any NA responses), whilst the overwhelming majority answer at least five out of ten questions. This leaves a small proportion (0.2%) who give N/A responses to more than half (five) of the ten questions. As discussed in the peer review, there is a case to be made for not scoring these respondents. However, it is instructive to look at which questions these respondents do and do not answer in order to inform this decision.

Table 4.4 shows the proportion of respondents responding to each item (i.e. giving responses other than NA) for those answering three, four, five and six questions out of ten.

Table 4.4					
Item	Question(s)	Overall base =3	Overall base =4	Overall base =5	Overall base =6
	No of respondents:	*5*	*10*	*38*	*80*
1	Chilling questions	100%	100%	100%	96%
2	Cooking food to steaming hot	0%	10%	26%	36%
3	Eating meat with pink/red juices	0%	30%	32%	59%
4	No of times would consider re-heating food	0%	0%	39%	64%
5	How usually tell food has been re-heated	0%	0%	24%	44%
6	Washing raw meat/poultry	0%	0%	3%	8%
7	Where/how store raw meat/poultry in the fridge	0%	40%	42%	50%
8	Washing hands	0%	20%	34%	44%
9	Knowledge and checking of use by dates	100%	100%	100%	100%
10	Last day would consider eating leftovers	100%	100%	100%	100%

None of the 15 respondents answering only three or four out of ten items give an answer for items 4, 5 or 6 (see shaded boxes in the table), whilst few of them answer items 2, 3, 7 or 8. This suggests that these respondents do little or no cooking or food preparation themselves and this would seem to provide a good justification for not scoring them, in addition to the proportion of items answered. The patterns are much less clear cut for those answering five or six out of ten items and it is hard to make an argument for not scoring these respondents without introducing more complex criteria.

RECOMMENDATION: Respondents answering less than half (five) of the ten items should not receive a score on the new IRP.

4.3.2 Patterns of DK responses

Table 4.5			
Item	Question(s)	Respondents giving DK responses	Percent
1	Chilling questions *	2,438	38.1%
8	Washing hands *	82	1.3%
5	How usually tell food has been re-heated	79	1.2%
7	Where/how store raw meat/poultry in the fridge *	26	0.4%
4	No of times would consider re-heating food	20	0.3%
9	Knowledge and checking of use by dates *	17	0.3%
3	Eating meat with pink/red juices	10	0.2%
6	Washing raw meat/poultry	6	0.1%
2	Cooking food to steaming hot	3	0.0%
10	Last day would consider eating leftovers	0	0.0%

* for these items comprising more than one question, DK responses on any of the relevant question are counted.

The number of DK responses per item is shown in Table 4.5. Only the chilling item has a substantial number of DK responses – most of these arise from those (2,351) respondents answering DK to Q4_12 about what temperature their fridge should be. Otherwise, there are fairly negligible numbers of respondents answering DK.

This is reflected in Table 4.6 which shows the distribution of the number of DK response across the ten items. Only a small proportion of respondents (2.1%) give two or more DK responses.

Table 4.6

Total DK responses	Number of respondents	Percent	Cumulative percent
0	3,871	60.5%	60.5%
1	2,386	37.3%	97.9%
2	120	1.9%	99.7%
3	13	0.2%	99.9%
4	4	0.1%	100.0%

It is difficult to assess the relative impact of DK responses in individuals' IRP scores without making reference to the total number of questions answered (i.e. the overall base). Table 4.7 shows the relationship between the number of DK responses and the overall base.

Table 4.7

Overall base	Total DK responses					
	0	1	2	3	4	Total
3	0	3	2	0	0	5
4	2	5	3	0	0	10
5	18	14	5	1	0	38
6	33	35	9	2	1	80
7	133	105	21	2	0	261
8	327	242	16	1	0	586
9	571	431	19	0	1	1022
10	2787	1551	45	7	2	4392
Total	3871	2386	120	13	4	6394

If respondents answer DK to a majority of questions there might be an argument for reconsidering the treatment of DK responses or perhaps for not scoring these respondents. However, this is not the case.

The small number of respondents giving three of four DK responses, answer at least five or six questions (respectively). For those giving one or two DKs, we have already recommended that the minimum overall base be set at five. Therefore, there would seem to be no justification for changing the scoring/treatment of DKs. On the other hand it would be useful to monitor this relationship in the future in case larger numbers of respondents answer DK (for example, if the survey were moved to a different mode).

RECOMMENDATION: Treat DK responses as per the current IRP but monitor the distribution of these answers in future waves.

4.3.3 Summary of changes to IRP scoring for missing data

The only change to the scoring for the new IRP is to introduce a minimum threshold of five answers out of ten items in order to receive a score. This change is likely to affect a very small number of respondents (less than ten).

4.4 Conclusion

The redeveloped IRP has undergone qualitative and quantitative review. Changes to the IRP have been made in response to these reviews and following a deliberative workshop. The IRP is now constructed using 16 questions from Food and You. These questions are used to measure 10 domestic food safety behaviours which are compiled in the IRP. With the changes made to the construction of the index following the review, we can have confidence that the IRP is a robust tool which can be used to explore domestic food safety practices and track changes in these practices over time.

Appendix A. Current Index of Recommended Practice

Food and You current composite measure – Technical Report

Aim
The aim was to redevelop the original Food and You composite measure of domestic food safety practices so that it could track progress towards the Agency's strategic priority to 'improve public awareness and use of messages about good food hygiene practice at home'.

Data used
Data from Wave 1 and Wave 2 of the Agency's Food and You survey was used to redevelop the composite. Fieldwork for Wave 1 of the survey was carried out in 2010, and fieldwork for Wave 2 of the survey was carried out in 2012[3].

How the current composite measure was calculated?
For each respondent in Wave 1 and Wave 2 of Food and You, a score was calculated based on the percentage of reported domestic food practices that were in line with Agency recommended practice. Scores took into account questions which were not applicable for respondents by removing these questions from that respondent's base.

In order to produce an overall performance indicator for the Agency, an average (mean) percentage of all the respondents' performance scores were calculated. This provides an average percentage of reported practices that were in line with Agency guidance.

Seventeen domestic food safety practices make up the composite measure. These questions are grouped into five separate categories: chilling, cooking, cleaning, cross contamination and use by dates. Separate performance indicators were also calculated for each of these five areas.

Questions used and response coding
In order to calculate scores for each respondent, individual question responses were coded as either recommended practice (RP) or non-recommended practice (NRP). Judgements about which responses should be coded as recommended and non-recommended practice were made in consultation with the FSA Hygiene and Microbiology Team. These are detailed, along with the questions which make up the composite, in the table below.

[3] For further information on Food and You and links to the reports see: http://www.food.gov.uk/science/research/ssres/foodandyou/

Food safety practice	RP response (1)	Non-RP response (0)	N/A
Chilling			
Q4.9 Do you ever check your fridge temperature?	- Yes - Someone else in the household does - I don't need to – it has an alarm	- No - Don't know	- N/A
Q4.10 How often do you or another person in your household check the temperature of the fridge?	- Daily - 2-3 times a week - Once a week - Less than once a week but more than once a month - Once a month - I don't need to – it has an alarm *If respondent said 'I don't need to – it has an alarm' in Q4.9, then coded RP in Q4.10*	- Four times a year - 1-2 times a year - Never - Don't know/ Can't remember	- If respondents said 'No' in Q4.9 - N/A
Q4.11 Thinking about fridge temperature, can you tell me how you normally check the temperature? *(multicode)*	- Check the temperature display /thermometer built into fridge' - Put a thermometer into the fridge and check *If respondent said 'I don't need to – it has an alarm' in Q4.9 and/or Q4.10, then coded RP in Q4.11*	- Check setting/gauge of fridge - Look inside/check for ice/condensation - Feel food inside to see if it is cold - Family/friend checks it for me - I do not check it - Other answer - Don't know	- If respondents said 'No' in Q4.9 - N/A
Q4.12 What do you think the temperature inside your fridge should be?	- 0-5°C	- Less than 0°C - More than 5°C but less than 8°C - 8-10°C - More than 10°C - Go by setting on the fridge - Other answer - Don't know	

Food safety practice	RP response (1)	Non-RP response (0)	N/A
Cooking and reheating			
Q4.1.13 Do you do the following things at all when you are in the kitchen and if so how frequently: - Cook food to steaming hot	- Always	- Most of the time - Sometimes - Never - Don't know	- N/A
Q4.1.14 Do you do the following things at all when you are in the kitchen and if so how frequently: - Eat chicken or turkey if the meat is pink or has pink or red juices	- Never	- Sometimes - Most of the time - Always - Don't know	- N/A
Q4.25 How many times would you consider re-heating food after it was cooked for the first time?	- Not at all - Once	- Twice - Three times - More than three times - Don't know	- N/A
Q4.26 And how do you usually tell that food has been re-heated properly? *(multicode)*	- Steam is coming out of it - Check the middle is hot - Use a thermometer/probe - When it is bubbling - When it is piping hot - Test with a knife/fork/spoon	- Taste it - Stir it - Check it is an even temperature throughout - Put hand over it/touch it - Use a timer - It looks hot - Experience/you just know - The smell of it - Check texture - Other answer - I don't check - Don't know	- N/A

Food safety practice	RP response (1)	Non-RP response (0)	N/A
Cross-contamination			
Q4.1.5 Do you do the following things at all when you are in the kitchen and if so how frequently: - Wash raw meat and poultry	- Never	- Sometimes - Most of the time - Always - Don't know	- N/A
Q4.3 Why do you think people wash chopping boards after preparing raw meat, poultry or fish? *(multicode)*	- To stop the remains of it getting onto the next food - It can be dangerous if you don't - To wash away germs/bacteria - To prevent food poisoning - To stop/prevent contamination/ cross contamination	- To stop the flavour/taste transferring to other foods - To get rid of the mess - As it looks dirty - It's a habit - Just what people do/are told to do - Don't know why - Other	- N/A
Q4.14 Where in the fridge do you store raw meat and poultry? *(multicode)*	- Bottom shelf - Separate compartment - Separate from any other foods - Separate/other fridge - Away from cooked meats	- Anywhere - At the top of the fridge - In the middle of the fridge - Wherever there is space - Put in a container in the fridge - Other Answer - Don't know	- Don't store raw meat/poultry in fridge - Don't buy/store meat/poultry at all - Kept in the freezer - N/A
Q4.15 How do you store raw meat and poultry in the fridge? *(multicode)*	- Away from cooked foods - Covered in film/foil - In a covered container - In a drawer/special compartment/allocated shelf in fridge - In plastic bags (any mention) - On a covered plate/bowl/dish	- In its packaging - On a plate - Covered with a plate/dish - Leave uncovered (any mention) - Other answer - Don't know	- Don't store raw meat/poultry in fridge - Keep in freezer - N/A

Food safety practice	RP response (1)	Non-RP response (0)	N/A
Cleaning			
Q4.1.11 Do you do the following things at all when you are in the kitchen and if so how frequently: - Wash hands before I start preparing or cooking food	- Always	- Most of the time - Sometimes - Never - Don't know	- N/A
Q4.1.12 Do you do the following things at all when you are in the kitchen and if so how frequently: - Wash hands after handling raw meat/fish	- Always	- Most of the time - Sometimes - Never - Don't know	- N/A
Use by dates			
Q4.19 Which of these indicates whether food is safe to eat? *(multicode)*	- Use by - It depends	- Best before date - Sell by date - Display until date - None of these - Don't know - All of these	- N/A
Q4.22 Do you check use by dates when you are about to cook or prepare food?	- Always - Depending on food type	- Sometimes - Never - Don't know	- N/A
Q4.24 If you made a meal on Sunday, what is the last day that you would consider eating the leftovers?	- The same day - Monday - Tuesday - Never have leftovers	- Wednesday - Thursday - Friday - Saturday - Sunday - More than a week - Don't know	- N/A

Appendix B. Overview of IRP peer reviews

Qualitative review of IRP

The qualitative review was undertaken by a food hygiene specialist, and addressed the following issues:

- The derivation of the Index of Recommended Practice
- The questions, scoring and weighting
- Measuring changes over time

An overview of the reviewer's comments for each of these areas is provided below, accompanied by FSA response where appropriate.

Index of Recommended Practice and its derivation

As part of the qualitative review, the derivation of the revised IRP was commented on. In considering the derivation of the IRP, the reviewer noted that the focus was on practices being in line with FSA advice, and commented that with the current construction adding weighting to scores will disrupt the accuracy of interpretation when NA responses are removed from the question base.

FSA response
The Agency agrees with the reviewer that weighting the questions will have implications for missing data (from 'not applicable' questions) as the overall score for respondents with missing data will be dependent on the pattern of questions which are applicable. While this is also true when weighting is not applied to the questions, weighting the questions would exacerbate this issue, meaning that overall scores of those with missing data would be less comparable to the scores of respondents who answer all questions. The Agency sees this as a reason to avoid adding weightings to the questions, as there is not a pragmatic way to overcome these issues.

The questions, scoring and weighting

The qualitative review considered each of the questions in the revised composite and how responses to the questions were defined in terms of whether the reported practices were in line with recommended practice or not. The review also considered whether the questions in the index should each be weighted according to their food safety importance / relevance.

The reviewer advocated weighting the questions individually in terms of estimates of the food safety 'risk' (probability and consequences) they each represent and the validity of the question. The reviewer suggested a weighting between 1 and 5 for each question.

FSA response
While the IRP would ideally be used as a measure of risk (and in this case it would be appropriate to weight questions) the index is unfortunately not able to provide a measure of risk due to the complex context from which risk arises, which is not possible to measure effectively using the questions in Food and You.

Questions related to chilling

The reviewer commented that these questions were relevant as growth of bacteria is temperature dependent. Many do not grow below 8°C but Listeria can grow down to 5°C. For the four questions on chilling, the reviewer proposed individual weights ranging from 1-5, with the lower weights representing that some of the questions are less useful on their own. One question asks about knowledge rather than reported behaviour. If questions relating to knowledge were to remain in the IRP it should be made clear in the definition that questions included cover both reported behaviour and knowledge.

Cooking and reheating

The second group of questions in the composite addresses practices relating to cooking and reheating. The reviewer commented that cooking and reheating is important as high temperatures kill bacteria and viruses. The suggested weights for these questions varied from 3-5, with lower suggested weights for questions where potential exceptions to RP were identified. For example, potential exceptions to RP were identified for the question relating to whether respondents always cook food to steaming hot throughout.

Cross-contamination

A further set of questions included in the index addressed practices relating to cross-contamination. The reviewer commented that cross-contamination is an important factor to be included as food which has been cooked, or is otherwise ready-to-eat, would be expected to be pathogen free. If this food becomes contaminated by uncooked food it may no longer be safe to eat. The reviewer suggested weighting ranging from 3-5 for individual questions relating to cross-contamination. Lower weights here were suggested for questions which were less useful alone (did not provide enough context), or where the reviewer identified potential exceptions to RP. While there are potential exceptions to RP (situations where recommended practice as defined by the composite may not be optimum or other practices equally appropriate), the FSA notes that Food and You is not sensitive enough to detect the contexts where these exceptions may occur.

Cleaning

The next two questions included in the composite address the practice of cleaning, and specifically hand washing. The reviewer commented that cleaning hands is particularly important as it is often not possible to prepare food without touching it, which may lead to the transfer of pathogens from the person handling food to the food or cross contamination (from other foods) via the hands of the person handling the food. The reviewer notes that there is a social desirability bias associated with cleaning, which may lead to over-reporting of this practice. Nevertheless a high weight (5) was applied to these questions to reflect their importance in food safety.

Use by dates

Three questions included in the index ask about use by dates and how long food is safe to eat for. The reviewer commented that use by dates are important as they are focused on food safety and set a period for which food is safe to eat, if stored correctly. If food is kept beyond the use by date there is a danger that pathogen population may grow to a dangerous size. The reviewer suggests reasonably high weights (4-5) for the questions on this topic.

High risk foods

The reviewer also considered whether there were any important elements of food safety which were not covered by the index. The reviewer identified that questions relating to specific high risk foods are not included in the index, but notes that this is incorporated elsewhere e.g. in questions relating to checking fridge temperature and 5°C being the recommended temperature as a preventative measure relating to *Listeria*.

Measuring changes in behaviour over time

The reviewer finally considered the ability of the index to measure changes in reported behaviour over time. He noted that this will depend on the degree to which scientific knowledge is likely to change. The core issues around high-risk food and cross-contamination are also unlikely to change and issues around raw and undercooked meat have also been established over a long period of time.

Quantitative review of the IRP

The quantitative review was carried out by a NatCen statistician, and is published in full. This review of the current IRP addressed three broad aspects of the composite measure:

1. Specification and validity
2. Construction and scoring
3. Sensitivity (to changes over time)

Two distinct types of construct

Latent variables measured with multi-item scales can take two distinct forms, they can be *reflectively*-indicated or *formatively*-indicated. The distinction is important as it affects the way in which a scale is evaluated.

Reflectively-indicated constructs are made up of items (questions) that *reflect* an underlying (latent) variable which is assumed to exist but cannot be measured directly e.g. general intelligence or social/political attitudes. These items typically have similar/related content and so tend be highly correlated with one another and therefore largely interchangeable. Formatively-indicated constructs on the other hand are made up of items which are multi-dimensional in nature but together form a composite which represents a recognisable collection of behaviours or concepts. The items which make up the scale (in fact the word index tends to be used rather than scale) may or may not be related and therefore are not necessarily interchangeable with one another.

The IRP is an example of a formatively-indicated construct: it is made up of five domains which together form a collection of behaviours that represent "food safety practice". A key issue with this kind of construct is that conventional procedures to assess validity and reliability of scales are not appropriate. In particular, measures of internal consistency such as Cronbach's alpha are not relevant and in fact could even be misleading or damaging. Moreover, little attention has been given in the literature as to how to evaluate such measures.

Specification and validity

If a formatively-indicated construct such as the IRP cannot be evaluated using conventional methods such as Cronbach's alpha then how should we go about assessing the validity of such measures? To answer this we have drawn upon a methodological paper by Diamantopolous and Winklhofer, *Index Construction with Formative Indicators: An Alternative to Scale Development*[4] which itself draws upon the dispersed literature on formative indicators to arrive at four criteria for successful index construction: content specification, indicator specification, indicator collinearity and external validity. We will look at each of these in turn:

[4] Adamantios Diamantopoulos A & Winklhofer H. Index Construction with Formative Indicators: An Alternative to Scale Development. Journal of Marketing Research Vol. 38, No. 2 (May, 2001)

1. Content specification – has the content been specified explicitly and in detail?

First, it is important to specify the scope or domain of content that the index is intended to capture. The current IRP is described as a measure of "domestic food safety practices" and is said to measure "the percentage of reported food safety practices that were in line with Agency recommended practice". It covers five domains of practice: chilling, cooking, cleaning, cross contamination plus use by dates.

2. Indicator specification - do the indicators (questions) cover the scope of the latent variable?

According to Diamantopolous and Winklhofer, the items used as indicators should cover the entire scope of the latent variable described in the content specification. Given the lack of an explicit definition it is difficult to assess the IRP under this criterion. However, it is worth noting that whilst the stated intention is to measure "food safety practice", 3 out of the 17 questions included measure knowledge rather than behaviour. These 3 items are therefore arguably outside of (or at best tangentially related to) the scope of the index if it intends to measure actual food safety practice rather than the knowledge which may or may not underpin an individual's practice.

The stated aim of the composite, however, is to "track progress towards the Agency's strategic priority to improve public **awareness** and **use of** messages about good food hygiene practice at home". The words in bold suggest that both knowledge (awareness) and practice (use of) are important to the FSA. The issue here is that the two have been conflated to some extent. To address this concern it is tentatively suggested that if knowledge questions are to be included they be used to evaluate whether a behaviour is in line with recommended practice rather than receiving scores in isolation.

3. Multicollinearity – is there evidence of excessive multicollinearity?

Unlike reflectively-indicated scales, composite measures are multi-dimensional in nature and therefore, as discussed, the items that comprise them may or may not be correlated. Moreover, excessive collinearity among the indicators is undesirable; in an extreme case the information from one indicator could be effectively redundant if it is mathematically described by one or more of the other indicators.

To investigate potential collinearity, tetrachoric correlation coefficients were calculated for all pairs of (dichotomous) RP variables that make up the composite measure.[5] These are shown in Appendix C. The highest correlation between any pair of items is between the two Cleaning questions (Q4_1_11 and Q4_1_12) demonstrating that if a respondent washes their hands before preparation/cooking they are more likely to wash their hands after handling raw meat/fish. However, whilst this relationship is fairly strong the two questions are far from "collinear".

What is most striking looking at the correlations is that a number of them are negative. This is not an issue for formatively-indicated constructs but it is instructive to investigate these correlations nevertheless. The largest negative correlation is between Q4_1_11 and Q4_1_5: the more likely you are to wash your hands before you start preparation/cooking the more likely you are to also wash raw meat/fish. This suggests that behaviours can be

[5] Pearson correlation coefficients are not appropriate for dichotomous variables. Instead tetrachoric correlation coefficients can be used for estimating the correlation under the assumption that the underlying variables are normally distributed

inconsistent in terms of safety and are therefore not necessarily rational or based upon sound knowledge/understanding but rather may be habitual and based upon a combination of factors including what may have been learnt from others. Qualitative research would be helpful here in drawing out the factors that underpin behaviour. This kind of insight could help to inform the design of instruments such as the composite measure.

4. External validity

Diamantopolous and Winklhofer make a number of suggestions for assessing external validity. Many of these suggestions involve building structural equation models to link the items and the overall composite to various external measures. Unfortunately this is somewhat outside the scope of this enquiry. A further suggestion is to correlate the index with other constructs with which it would be expected to be linked (i.e. antecedents and/or consequences). An obvious consequence of (poor) food safety practice is food poisoning, although this can obviously occur outside of the home and be nothing to do with practice in the home. However, for what it's worth Table B1 shows the overall score for those who said they had and hadn't had food poisoning. It shows that the score for those who had had food poisoning is (significantly) higher than the score for those that hadn't. If food poisoning is a consequence of bad practice then we might expect the relationship to be in reverse. However, there are many factors here such as non-domestic food practices, differences in individual susceptibility, and recollection bias. There is also the issue of time: practice may have changed over time but it is only measured in the present whilst food poisoning is measured looking backwards. It may well be that having had food poisoning makes one more careful but realistically it is virtually impossible to conclude anything from this relationship.

Table B1	
Had food poisoning?	Overall IRP
Yes	0.741
No	0.728
Total	0.733

Summary

In summary, whilst the external validity of the index is rather difficult to assess, there are clearly some issues with the definition/scope of the index which need to be addressed.

FSA response

The reviewer details the description of the index, and suggests that the definition is not explicit. The Agency acknowledges this point, and in future will aim to ensure that this and similar constructs are specified appropriately. The Agency also notes the contradiction between the intention and aim of the composite, and aims to ensure that this is clarified by further considering the role of knowledge questions in the composite.

The reviewer summarises that the external validity of the index is difficult to assess, and there are issues with the definition / scope of the index which should be addressed. The agency is keen to address these issues by clarifying the definition of the index.

Construction and scoring of the current IRP

1. Has the IRP been constructed as specified?

The composite measure has been constructed as specified in Appendix A.[6]

2. Is the scoring of the overall composite fair and logical?

To answer this we should consider the relative weight given to each question and the relative weight given to each domain. We should also consider how missing data is handled and any implications this might have.

The overall score is constructed thus:
OverallRP = Overalltotal / Overallbase

In other words the total number of questions scored as "following recommended practice" (Overalltotal) divided by the total number of questions answered (Overallbase). The questions answered therefore receive equal weight in the IRP. This is parsimonious and avoids any difficult value judgements about the relative importance of different aspects of food safety practice, which would always be open to question. However, there are two important consequences of scoring in this way when considering the fact that the overall composite is also split into five domains.

The first is that, as there are different numbers of questions in each domain, the domains (effectively) receive different weights in the overall score. For example, for respondents who answer all 17 questions in the composite, Cross-contamination receives twice as much 'weight' as Cleaning due to having four questions rather than two. This may or may not be desirable but is certainly open to question.

Second is the fact that different respondents can receive different weights for the same domain due to the way that missing data is handled.[7] (This issue is explored in more detail below but here we consider the way in which it affects the relative weighting of domain scores for different respondents.) For example, a respondent who answers all 17 questions will (effectively) receive a weight of 2/17 (0.117) for Cleaning (in other words Cleaning will make up 11.7% of their overall score) whilst a respondent answering all questions except one of the Cleaning questions will receive a weight of 1/16 (0.0625) (in other words Cleaning will make up 6.25% of their overall score). A further consequence of this is that two individual respondents can receive the same scores on all five domains but different overall scores. Inspection of the data reveals that this does occur; in one example two respondents with the same domain scores receive quite different overall scores, 0.5 and 0.64, due to different patterns of non-response. There is a clear logical inconsistency here no matter how you view the fairness of the scoring per se.

The extent to which these issues are seen as problematic depends to some degree on the way the overall composite is viewed in terms of the relationship between the questions, the domains and the overall score. If the overall composite is viewed as being constructed from five domains which in turn are constructed from individual questions, then the differential

[6] One respondent scored 1.75 on the Cooking domain. This has been rectified for the purposes of the analysis.

[7] Missing data refers to 'NA' answers and other missing data. It does not refer to 'don't know' as these are not treated as missing but allocated a score of zero.

weighting of domains is certainly open to question. In this case consideration should be given to domains receiving equal weighting in the overall composite.

3. Is the scoring of the five domains fair and logical?

There are three aspects to consider here: the relative weighting given to different questions within domains, any dependency in the scoring within domains (i.e. the score on one question depends on the outcome of another question) and the way in which missing data is handled. In this section, we explore the first two issues, whilst the last, that of missing data, is discussed in detail below.

In the IRP, equal weight is given to questions within each domain. Again, this is parsimonious and avoids difficult value judgements about the relative importance of different aspects of practice.

Unlike the other four domains, the Chilling domain includes some dependency in the scoring which makes it a little more complicated in its construction and deserving of scrutiny. There are four questions that make up this domain. However, only if the answer to Q4.9 (Do you ever check your fridge temperature) is Yes, do respondents get scored on all four questions; if the answer is no then the following two questions about how this is done are not applicable and therefore not scored (and the base reduces from 4 to 2).

To evaluate the effect of this scoring system, it is instructive to look at all the possible combinations of answers to the four questions and the scores that these lead to; these are shown in Table B2.

Table B2

Combination	Score	Check temperature	Check often enough	Check in right way	Know right temperature
1	0.00	0	-	-	0
2	0.25	1	0	0	0
3	0.50	0	-	-	1
4	0.50	1	1	0	0
5	0.50	1	0	1	0
6	0.50	1	0	0	1
7	0.75	1	1	1	0
8	0.75	1	1	0	1
9	0.75	1	0	1	1
10	1.00	1	1	1	1

From this we can see that it is possible to score 0.5 on the Chilling domain simply by knowing the correct fridge temperature despite admitting to not ever checking it (combination 3). In this case practice is absent altogether. Moreover, the same score is given for three other combinations (4-6) all involving checking the temperature.

It is also possible to score 0.75 despite not checking in the right way (combination 8). The same score can be had for two other combinations (7 and 9) which both involve checking in the right way and therefore, arguably, should command higher scores than combination 8.

There are some clear logical inconsistencies here and the scoring of this domain needs reviewing.

4. Has missing data been dealt with appropriately?

An important consideration in the construction of any scale or index using survey data is how missing data is handled. We have already seen that the way that missing data is handled has important implications when respondents have different patterns of non-response across domains. However, it is also important to consider the overall number/proportion of questions that are answered by individual respondents both within each domain and across all five domains. For example, if a respondent were to answer only one or two questions out of the 17 that make up the composite, should they receive an overall score? Similarly if they only answer 1 question out of 4 for Cross-contamination should they receive a score for this domain?

There are no hard and fast rules for dealing with missing data but a common approach with reflectively-indicated constructs is to only give a score to respondents who have answered a reasonable proportion of questions, with the threshold for inclusion set with reference to the trade-off between the number of questions answered and the proportion of respondents who would be excluded (not scored). However, in reflectively-indicated constructs indicators are interchangeable and therefore if some are missing it doesn't change the meaning of the composite. For a formatively-indicated construct such as the IRP, on the other hand, missing data may represent a "hole" in the latent variable that the composite is purporting to measure and it is therefore of more concern, particularly if data is missing altogether from a batch of questions which either implicitly or explicitly constitute a sub-domain of the composite. For example, in the case of the IRP, if a respondent has no data on either of the Cleaning questions they won't receive a score for the Cleaning domain and it is questionable therefore whether the overall score for this respondent is comparable with the overall score for a respondent who does receive a Cleaning score.

So we see it is important to consider three aspects: (i) the number/proportion of questions answered overall; (ii) the number answered within each domain; and (iii) the pattern of response across domains.

Looking first at the number of questions answered overall, the minimum Overallbase in waves 1 and 2 is 6 (about 1 in 3 questions answered) with 0.3% of the sample answering fewer than 9 (about 1/2). There is a strong argument for not scoring these individuals given the small numbers involved and the fact that they failed to answer the majority of questions.

However, as we have argued, it is also important to look at the pattern across domains. Analysis of waves 1 and 2 combined shows that a total of 168 respondents (2.6% of the sample) did not receive a score on at least one domain: 11 respondents (0.2%) received a score for 3 out of 5 domains, whilst a 157 respondents scored 4 out of 5 (2.5%). The most common domain to be missing was Cleaning (129 respondents in total). Again there is an argument for not giving these respondents an overall score.

Turning to missing data within domains, it is worth considering the fact that each domain is a formatively-indicated construct (as is the overall composite) and therefore as we argued above, the questions that make up a domain are not interchangeable. So the same argument applies about "holes" in the latent variable, although arguably perhaps not to the same degree given the narrower focus of the domains. Nevertheless, if a respondent answers only one question in a domain comprising 3 or 4 items it is questionable whether their score is comparable with a respondent who answers all 3 or 4 questions, particularly as they will by definition score either 1 or 0. Currently the scoring system is based on the number of questions answered and one question is enough to receive a score on a domain. There is clearly an argument for increasing the minimum threshold although in doing so the trade-off, in terms of losing too many respondents, must be considered.

Summary

In summary there are a number of aspects to the current scoring system that are open to question. Specifically it is recommended that the following be reviewed and possibly revised:

- weighting of the overall composite measure
- the way in which missing data is handled
- the scoring of the Chilling domain

FSA response

Although the Agency takes on board the reviewer's comments about 'holes' in the index when questions are not answered, the Agency does not think this is problematic as it results in the index providing a score which relates to the food safety practices that each individual undertakes. For example, the question 'how frequently do you eat chicken or turkey if the meat is pink or has pink / red juices' is unlikely to be applicable for a vegetarian. As this is not a behaviour that they undertake it does not contribute to their overall 'food safety practices'. As such, it makes sense to remove this question from their questions base. In essence, 'food safety practices' are variable and honed to individuals, and so removing questions from the question base does not necessarily represent a 'hole' in the latent variable.

Sensitivity

Power to detect changes in the current IRP over time

Finally we consider the ability of the current index to detect changes over time. Assuming that the sample size is consistent over time the survey will continue to be able to detect fairly small changes in food safety practice as measured by the IRP. Changes of around 1.2 points out of 100 (equivalent to half a standard deviation) are detectable at the 95% significance level (with 80% power). In other words, if the mean of the index increases from74.0 to 75.2 in the population as a whole, the survey would have a good chance of detecting this difference. The increase in the IRP from Wave 1 to Wave 2, however, despite being statistically significant, might or might not reflect a real change in practice given the scoring issues identified above.

Implication of changing scoring from binary to ordinal

The question of whether changing the scoring from binary (1=RP;0=Not RP) to ordinal (e.g. utilising a scale from "always follows RP" down to "never follows RP" with gradations in between) would increase the ability of the index to measure changes over time is very difficult to assess. It is self-evident that using an ordinal scale, other things remaining equal, would increase the sensitivity of a single measure to *actual* changes in practice and therefore would increase the sensitivity of a composite measure based on a collection of such individual measures.

However, quantifying the degree of additional sensitivity would only be possible if an explicit set of assumptions were made about how an improvement in practice would translate into

movement along such a scale (e.g. the percentage of people who might move from “never” to “sometimes”, from “sometimes” to “usually” and from “usually” to “always”, say). Furthermore, changing the way in which the index is scored would change its meaning and this would also have to be taken into account.

There is also the question of how to attach scores to points along a scale. For example, whilst the scale itself might be ordinal, there might be a strong theoretical (food safety) argument for the scoring to reflect the relative safety/harm of various answers. For example, it might be argued that a fridge temperature of 5-8 degrees should receive a score closer to 1 than 0 (for reasons already outlined in earlier sections); however, it is not clear how any such score should be determined except via some kind of subjective judgement of the kind which it would ideally be prudent to avoid.

Appendix C. Tetrachoric correlations

The strength of association between two variables can be summarised with a correlation coefficient. A general rule of thumb for interpreting these coefficients is that a score of:

- 0.50 or more indicates a strong association between two variables (dark shading in the table below);
- 0.30 to 0.49 indicates a weak to moderate association (light shading); and
- Less than 0.30 indicates little, or no, association (not shaded).

	Check fridge temperature	Know fridge temperature	Cook food to steaming hot	Eat if pink/ has pink juices	Re-heat food	Check re-heated food	Wash hands preparing or cooking food	Wash hands after handling meat/fish	Wash raw meat and poultry	Wash chopping boards	Where store in fridge	How store in fridge	Indicates safe to eat	Check use by dates	Eating leftovers
Check fridge temperature	1.00														
Know fridge temperature	0.24	1.00													
Cook food to steaming hot	0.03	-0.01	1.00												
Eat if pink/ has pink juices	0.00	0.00	0.32	1.00											
Re-heat food	0.04	0.00	0.20	0.23	1.00										
Check re-heated food	0.08	0.14	0.16	0.11	0.08	1.00									
Wash hands preparing or cooking food	0.16	-0.03	0.33	0.21	0.21	0.08	1.00								
Wash hands after handling raw meat/fish	0.17	0.02	0.39	0.31	0.20	0.12	0.62	1.00							

Wash raw meat and poultry	-0.11	0.07	-0.09	0.05	0.00	0.06	-0.22	-0.18	1.00						
Wash chopping boards	0.10	0.13	0.11	0.21	0.12	0.20	0.12	0.23	-0.04	1.00					
Where store in fridge	0.14	0.15	0.14	0.12	0.21	0.19	0.20	0.24	-0.07	0.21	1.00				
How store in fridge	0.10	0.00	0.04	0.11	0.02	0.06	0.12	0.13	-0.17	0.07	0.15	1.00			
Indicates safe to eat	0.00	0.11	-0.01	0.02	0.04	0.09	-0.03	0.02	0.09	0.13	0.10	-0.01	1.00		
Check use by dates	0.20	0.14	0.19	0.13	0.23	0.17	0.28	0.31	-0.08	0.20	0.18	0.03	0.21	1.00	
Eating leftovers	0.05	-0.14	0.17	0.10	0.36	-0.01	0.21	0.22	-0.15	-0.06	0.06	-0.02	-0.14	0.25	1.00

(obs=6369)

Appendix D: New revised Index of Recommended Practice

Item	Question	RP response (1)	Non-RP response (0)	N/A	Final item scoring for combined items
Chilling	Q4.9 Do you ever check your fridge temperature?	-Yes - Someone else in the household does - I don't need to – it has an alarm	- No - Don't know	- N/A	1 = RP responses to all questions 0 = Non-RP responses to 1 or more questions NA = NA to Q4.9
	Q4.10 How often do you or another person in your household check the temperature of the fridge?	- Daily - 2-3 times a week - Once a week - Less than once a week but more than once a month - Once a month - I don't need to – it has an alarm *If respondent said 'I don't need to – it has an alarm' in Q4.9, then coded RP in Q4.10*	- Four times a year - 1-2 times a year - Never - Don't know/ Can't remember	- If respondents said 'No' in Q4.9 -N/A	
	Q4.11 Thinking about fridge temperature, can you tell me how you normally check the temperature? *(multicode)*	AT LEAST ONE OF: - Check the temperature display /thermometer built into fridge' - Put a thermometer into the fridge and check *If respondent said 'I don't need to – it has an alarm' in Q4.9 and/or Q4.10, then coded RP in Q4.11*	ZERO RP RESPONSES AND AT LEAST ONE OF: - Check setting/gauge of fridge - Look inside/check for ice/condensation - Feel food inside to see if it is cold - Family/friend checks it for me - I do not check it - Other answer - Don't know	- If respondents said 'No' in Q4.9 - N/A	

	Q4.12 What do you think the temperature inside your fridge should be?	- 0-5°C	- Less than 0°C - More than 5°C but less than 8°C - 8-10°C - More than 10°C - Go by setting on the fridge - Other answer - Don't know		
Cooking food to steaming hot	Q4.1.13 Do you do the following things at all when you are in the kitchen and if so how frequently: - Cook food to steaming hot	- Always	- Most of the time - Sometimes - Never - Don't know	- N/A	-
Eating chicken/turkey if meat is pink or has pink/red juices	Q4.1.14 Do you do the following things at all when you are in the kitchen and if so how frequently: - Eat chicken or turkey if the meat is pink or has pink or red juices	- Never	- Sometimes - Most of the time - Always - Don't know	- N/A	-
Number of times you would consider re-heating food	Q4.25 How many times would you consider re-heating food after it was cooked for the first time?	- Not at all - Once	- Twice - Three times - More than three times - Don't know	- N/A	-

How you usually tell food has been re-heated properly	Q4.26 And how do you usually tell that food has been re-heated properly? *(multicode)*	AT LEAST ONE OF: - Steam is coming out of it - Check the middle is hot - Use a thermometer/probe - When it is bubbling - When it is piping hot - Test with a knife/fork/spoon	ZERO RP RESPONSES AND AT LEAST ONE OF: - Taste it - Stir it - Check it is an even temperature throughout - Put hand over it/touch it - Use a timer - It looks hot - Experience/you just know - The smell of it - Check texture - Other answer - I don't check - Don't know	- N/A	-
Washing raw meat/poultry	Q4.1.5 Do you do the following things at all when you are in the kitchen and if so how frequently: - Wash raw meat and poultry	- Never	- Sometimes - Most of the time - Always - Don't know	- N/A	-
Where/how you store raw meat and poultry in the fridge	Q4.14 Where in the fridge do you store raw meat and poultry? *(multicode)*	AT LEAST ONE OF: - Bottom shelf - Separate compartment - Separate from any other foods - Separate/other fridge - Away from cooked meats	ZERO RP RESPONSES AND AT LEAST ONE OF: - Anywhere - At the top of the fridge - In the middle of the fridge - Wherever there is space - Put in a container in the fridge - Other Answer - Don't know	- Don't store raw meat/poultry in fridge - Don't buy/store meat/poultry at all - Kept in the freezer (ONLY) - N/A	1 = RP responses to both questions 0 = Non-RP response to one or both questions AND no NA responses NA = NA response to one or both questions
	Q4.15 How do you store raw meat and poultry in the fridge? *(multicode)*	AT LEAST ONE OF: - Away from cooked foods - Covered in film/foil - In a covered container - In a drawer/special compartment/allocated shelf in fridge - In plastic bags (any mention) - On a covered plate/bowl/dish	ZERO RP RESPONSES AND AT LEAST ONE OF: - In its packaging - On a plate - Covered with a plate/dish - Leave uncovered (any mention) - Other answer - Don't know	- Don't store raw meat/poultry in fridge - Keep in freezer (ONLY) - N/A	

Washing hands before food preparation/after handling raw meat/fish	Q4.1.11 Do you do the following things at all when you are in the kitchen and if so how frequently: - Wash hands before I start preparing or cooking food	- Always	- Most of the time - Sometimes - Never - Don't know	- N/A	1 = RP responses to both questions 0 = Non-RP response to one or both questions NA = NA response to both questions
	Q4.1.12 Do you do the following things at all when you are in the kitchen and if so how frequently: - Wash hands after handling raw meat/fish	- Always	- Most of the time - Sometimes - Never - Don't know	- N/A	
Knowledge and checking of use by dates	Q4.19 Which of these indicates whether food is safe to eat? *(multicode)*	AT LEAST ONE OF: - Use by date - It depends	ZERO RP RESPONSES AND AT LEAST ONE OF: - Best before date - Sell by date - Display until date - None of these - Don't know - All of these	- N/A	1 = RP responses to both questions 0 = Non-RP response to one or both questions AND no NA responses NA = NA response to one or both questions
	Q4.22 Do you check use by dates when you are about to cook or prepare food?	- Always - Depending on food type	- Sometimes - Never - Don't know	- N/A	

Last day you would consider eating Sunday leftovers	Q4.24 If you made a meal on Sunday, what is the last day that you would consider eating the leftovers?	- The same day - Monday - Tuesday - Never have leftovers	- Wednesday - Thursday - Friday - Saturday - Sunday - More than a week - Don't know	- N/A	-