I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16137

Nudging in Complex Environments

Alexander K. Koch
Dan Monster
Julia Nafziger

MAY 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16137

Nudging in Complex Environments

Alexander K. Koch
Aarhus University and IZA

Dan Monster
Aarhus University

Julia Nafziger
Aarhus University and CEPR

MAY 2023

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Nudging in Complex Environments*

To study the effects of reminder nudges in complex environments, we apply a novel experimental approach based on a computer game in which decision makers have to pay attention to and perform multiple actions within a short period of time. The set-up allows us, first, to test the effect of reminders both on reminded and non-reminded actions and thus to observe whether reminders have (positive or negative) spillovers. Second, we investigate spillovers between multiple nudges by testing the effect of scaling up the number of reminded actions. Third, we study intertemporal spillovers by investigating whether the effects of having been exposed to reminders persist after reminders are withdrawn. We observe that reminders have positive effects in the short run – multiple reminders more so than single reminders: while reminders lead to crowding-out of non-reminded actions, the positive effect on the reminded actions dominates. Yet, after withdrawal of the reminders, the negative spillover effect persists, while the positive effect partially fades out so that, overall, reminders have no effect.

JEL Classification: C9, D91

Keywords: nudging, spillover effects, attention, reminders, persistence, game-based experiments

Corresponding author:
Julia Nafziger
Department of Economics and Business Economics
Aarhus University
Nordre Ringgade 1
8000 Aarhus C
Denmark

E-mail: jnafziger@econ.au.dk

* The study is pre-registered on the AEA RCT Registry on the 21/11/21; the amendment on the 23/8/22 pre-registers the additional treatments without feedback (AEARCTR-0007932, https://www. socialscienceregistry.org/trials/7932/history). We would like to thank Jacob Sherson and his team for programming the game. For their comments we would like to thank Christoph Engel, Jenny Kragl, Andreas Roider, seminar/workshop participants in Aarhus, Bonn, Helsinki and London, and participants at the workshop Bounded Rationality: Theory and Experiments in Tel Aviv, the ESA meeting in Bologna, and the Verein für Socialpolitik meeting in Basel. This project is part of the SafeConsume project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727580.

1 Introduction

Individuals often act in complex environments where limited attention and focus may prevent them from taking the actions that are in their own best interest. Nudging is a popular approach to steer the behavior of individuals in a desired direction (Thaler and Sunstein, 2008). Yet, many studies, and especially studies on reminder nudges, focus on nudges to take *one* specific action or to pay attention to *one* specific factor, such as booking an appointment at the dentist, getting a flu vaccination, or paying a fee in time (for examples of nudges employed in recent studies, see Hummel and Maedche, 2019; DellaVigna and Linos, 2022).

Focusing on single behaviors however ignores the complexity of most real-world environments, where a decision maker has to pay attention to and take several actions within a short period of time. In such an environment, a nudge on one action may have positive or negative spillover effects on other actions which may reinforce or reverse the overall effectiveness of the nudge. Yet, testing the effectiveness of a nudge in a complex environment is challenging because it is difficult to unobtrusively observe a wide range of behaviors.

The aim of this paper is to study the effects of reminder nudges in complex environments using a novel experimental approach based on an online computer game that allows us to observe a wide range of behaviors. The game exposes subjects to a complex environment in which they need to pay attention to and perform multiple actions within a short period of time. Besides measuring the effect of a reminder nudge on reminded actions, our setup allows us to test for three types of (positive or negative) spillovers that such a nudge might have. First, we can identify the impact of reminder nudges on non-reminded actions. Second, we can investigate spillovers between multiple nudges by testing whether scaling up the number of reminded actions helps subjects make better decisions in the targeted and non-targeted domains or leads to (attention) overload. Third, we can study intertemporal spillovers by investigating whether the effects of having been exposed to reminders persist after reminders are withdrawn.

The game is on domestic food preparation. We use this context for two reasons – noting that it has generic features of complex environments that apply to many other contexts. First, domestic food preparation lends itself to studying the role of nudges that can help focus attention on correct behavior, because inappropriate food handling practices are widespread and the cause of food borne illness.[1] Second, domestic food preparation is characterized by complexity. People need to know about and apply a range of actions to prevent food borne illness, for example, by avoiding cross-contamination. Though, when cooking, they may have other things than food safety at the top of their mind (e.g., the recipe), be time constrained, and face disturbances (e.g., text messages, children, or pets). Lack of attention means that

[1] 9.4 million cases of food borne illness, 55,961 hospitalizations, and 1,351 deaths occur each year in the U.S. (Scallan et al., 2011, cf.). 10–30% of the cases of food borne illness can be attributed to domestic food preparation (for the US and Europe, respectively, see Dewey-Mattia et al., 2018; EFSA and ECDC, 2018).

knowing about correct behavior does not necessarily lead to correct behavior.[2]

Correspondingly, in our game, the decision maker has to prepare a dish under time pressure and faces disturbances while cooking. In the process, the decision maker can take a range of food safety related actions, such as washing hands, rinsing vegetables, checking whether the meat is done, as well as cleaning surfaces and kitchen utensils.

To test the effects of nudges, we implement three different treatments in a between-subjects design on the Prolific crowd working platform with more than 1500 subjects (see Figure 2 in Section 2). In *Control*, subjects simply play the game. In *Reminder*, subjects play the game as in *Control* but receive reminders about one type of food safety action (washing hands) at exogenously fixed points in time. In *ManyReminders*, subjects receive reminders about three types of food safety actions (washing hands, cleaning surfaces, and checking the temperature of the meat). Two days after the intervention, subjects play the game again, but now the reminders are withdrawn in all treatments.

In line with the previous literature on reminders, the comparison of *Reminder* with *Control* shows that a reminder has an immediate direct positive effect on the reminded action. By bringing an action to the top of the mind of the decision maker, a reminder increases attention to the focal action, and in doing so, prompts action. The contribution of our paper is to place this effect in the overall context of a complex environment and to provide evidence on the variety of spillovers – both positive and negative – that may arise in such environments.

First, we find that reminders undermine performance in non-targeted behaviors. Overall performance however still improves, because the direct impact on the reminded action is larger than the negative spillover effects on the non-reminded actions. These may arise, for example, because reminders distract attention from the other actions or because they (or the induced action) increase the costs of taking other actions. Still, the absence of any positive spillovers appears surprising in our setting because one could have expected that anticipated reminders help to freeze attention to other actions; and because the similarities of some actions might help, when seeing a reminder on one action, to recall and subsequently perform a similar action.

Second, the comparison between *ManyReminders* and *Reminder* shows that reminder nudges can be scaled in the sense that increasing the number of reminded categories leads to an improvement in overall performance. Yet, again, this has negative spillover effects on non-targeted behaviors. We also find that reminders have decreasing returns to scale because receiving several reminders at the same point in time is less effective than receiving a single reminder – indicating that paying attention to reminders is costly.

Third, we document intertemporal spillovers to behavior when the nudge no longer is present

[2] Gabaix (2019), for example, notes "[...] if people show ignorance in a survey, it is good evidence that they are inattentive. However, when they exhibit knowledge, it does not mean that they actually take into account the variable in their decision."

in game play two days after the intervention. Repeatedly receiving reminders and performing a behavior may help to build routines that persist even without reminders. In line with this, having been treated with reminders leads to better performance on the targeted behavior even after withdrawal of the reminders – however the effect is smaller than that of contemporaneous reminders. Yet, the negative spillover effect on other behaviors does not diminish in size. As a result, treatment with reminders has no net effect on overall performance after reminders are withdrawn.

In terms of mechanisms, our data suggest that reminders to some extent work mechanically by bringing an action to the top of the mind and therefore only partially succeed in building routines. Further, even before their withdrawal, reminders in *ManyReminders* become less effective over time, while a single reminder does not lose its effectiveness. These findings suggest that (too) many reminders may induce attention overload or be an annoyance, which also makes them less effective at building persistent behavior.

Finally, when exploring a number of alternatives to attention as potential mechanisms, we find no evidence (i) for spillovers from reminders arising because of licensing (an individual compensates for 'good' behavior induced by reminders with 'bad' behavior on other actions); (ii) for reminders to work because they signal to subjects what the 'desired' actions are; (iii) for reminders to work by inducing better knowledge. Further, while subjects react to performance incentives, subjects do not seem to respond to relative incentives in the time/effort costs of taking one action relative to another. Subjects do however react to visual cues that act as attention grabbers – suggesting that attention rather than time/effort considerations matter for spillovers.

Related literature

Only few studies consider the effects of nudges on different actions, tasks or factors and the spillover effects of nudges. Most closely related to our work are Altmann et al. (2022), Medina (2021), and Trachtman (2022) with their focus on cognitive capacity limitations and limited attention.[3] The novel features of our study is that we study a complex environment and that one can expect both positive and negative spillovers in our setting.

Neisser and Becklen (1975) (and follow-up studies in the area of visual attention), as well as Altmann et al. (2022), differ from our work in that their tasks are chosen to be cognitively demanding so that performing in one of the tasks directly impacts the mental or effort costs of the other task. Consequently, only negative spillovers can be expected. Altmann et al. (2022)

[3] Another strand of the literature analyzes spillover effects of nudges with a focus on preference or information channels (see Tiefenbeck et al., 2013; Brandon et al., 2019; Castro et al., 2020; Jessoe et al., 2021; Sherif, 2021; Hussam and Oh, 2022). Further, several studies test for spillover effects of behaviors, where the behavioral change however is not induced by a nudge (e.g., Lacetera et al., 2012; Reinstein, 2011; Cairns and Slonim, 2011; Null, 2011; Meer, 2017; Filiz-Ozbay and Uler, 2019; Ek, 2017, 2018; Ek and Miliute-Plepiene, 2018; Seres et al., 2021). See also Dolan and Galizzi (2015) for an overview of different kinds of behavioral spillover effects.

show that policies that promote active choice in one task, or make a task easily accessible, increase active choice in this task (adding numbers), but decrease performance in another task (memorizing numbers). Medina (2021) shows that reminding people to repay their credit card debt reduces late repayments, but increases overdraft fees for checking accounts. Unlike us, these papers do not study the effects of multiple nudges or test for spillovers after the withdrawal of the nudge. Trachtman (2022) is closest to our study in that she studies treatments with one and two reminders on two tasks. In comparison to our study, the tasks are unrelated, not temporally connected, and the environment is not complex. Using an app, she sends out messages about either taking action x (meditating) or taking y (tracking meals), or both on a given day. She finds positive direct effects and negative spillover effects that also persist after the withdrawal of the reminder. Yet, receiving both messages does not significantly change the direct or spillover effects.

We also contribute to the literature that addresses whether nudges can help build persistent behavior. In field experiments, long-term effects of nudges are often addressed (see, e.g., Allcott and Rogers, 2014; Calzolari and Nardotto, 2017; Sunstein, 2017; Frey and Rogers, 2014; Byrne et al., 2022). Lab experiments rarely study persistent behavior. An exception is de Haan and Jona (2018), who demonstrate spillover effects of nudgdes on future decisions – an observation that they explain with reinforcement learning.

Finally, we contribute to the variety of studies demonstrating that nudges can have negative effects, such as negative general equilibrium effects (Spiegler, 2015; Kamenica et al., 2011; Grubb, 2014; Grubb and Osborne, 2014; Duarte and Hastings, 2012; Handel, 2013), psychic costs (e.g., Allcott and Kessler, 2018; Jimenez-Gomez, 2018; Thunström et al., 2018; Thunström, 2019; Farhi and Gabaix, 2020), unanticipated effects (e.g., Damgaard and Gravert, 2018; Taylor, 2020; Hall and Madsen, 2022), shifts of attention (Caplin and Martin, 2017; Nafziger, 2020), reduction of learning in society (Carlin et al., 2013), substitution for habit formation (Taubinsky, 2013) or procrastination (Ericson, 2017).

2 Experimental design and procedures

We recruited UK residents using the Prolific crowd working platform for an online experiment over two nonconsecutive days. Subjects were randomized into three different conditions (2 treatments, 1 control). Our main study (called the *feedback study*) ran from November-December 2021. In November 2022, we conducted a second study (called the *no feedback study*), which differed from the main study in that subjects did not receive any feedback about their performance in the game. In the following, we describe the design of the main study and subsequently explain the amendment for the second study.

2.1 The online game

Game environment. Subjects play through a series of modules of a computer game that is about preparing food in a home kitchen setting (see here for a version of the game: https://safeconsume.eu/tools/safeconsume-game; see Figure 1 for some screenshots). The goal of the game is to prepare a dish under hygienic circumstances, subject to time pressure and disturbances. The game is set in a kitchen, which comprises a fridge, a sink (next to which there are placed soap, detergent, a brush, and an all purpose cleaner), a worktop with a cutting board and knife, a pan located on the stove, a food thermometer, and a rubbish bin. To complete a level, the subject has to prepare one or more recipes (depending on the module). Each recipe consists of three types of ingredients: chicken, a raw vegetable or fruit, and bread. Ingredients are stored in the fridge or the bread basket, and they need to be prepared by cutting them into smaller pieces. Only one food item at a time can be cut. The meat needs to be cooked. Once all food items are prepared and placed on a plate, a subject can press the 'serve' button. The recipe is completed by pressing the 'leave kitchen' button.

Performance. We measure performance based on a number of exogenously fixed critical handling points, to which we refer as **important food safety actions** (IFSAs). These are based on microbiological advice about safe food handling and encompass actions such as washing hands, cleaning kitchen utensils and surfaces, rinsing vegetables and checking the temperature of the meat (for details see Koch et al., 2022). IFSAs are listed in Table 1.

Added complexity. When cooking, people often face time pressure and disturbances. Applying the correct food safety actions therefore is not only about knowing them, but also about paying attention to them. To mirror such real life situations, we implement two measures that increase the complexity of the environment. First, a clock is running during the game. The game does not stop when time runs out, much like individuals would be able to continue meal preparation in real life if it took longer than planned.[4] Instead, after 5 minutes the timer turns red and subjects are incentivized to finish within the time limit. Second, in some recipes a cat appears that makes annoying sounds and leaves hair on the worktop. While the cat does not affect the cooking process, it is a disturbance. The longer the cat stays, the more hair it leaves. Subjects can remove the cat by clicking on it and moving it to a target field at the bottom of the screen (see the bottom right panel of Figure 1). But there is no reward for doing so.

Feedback stage. In the main parts of the experiment, a level consists of four recipes and each level ends with feedback about the performance in the game. Subjects are shown a screen with the overall IFSA score and subsequent screens listing how many times each IFSA was

[4] A time-out would endogenously determine how many food safety actions a decision maker could apply and would create problems for the empirical identification strategy.

Figure 1: Screenshots of the game

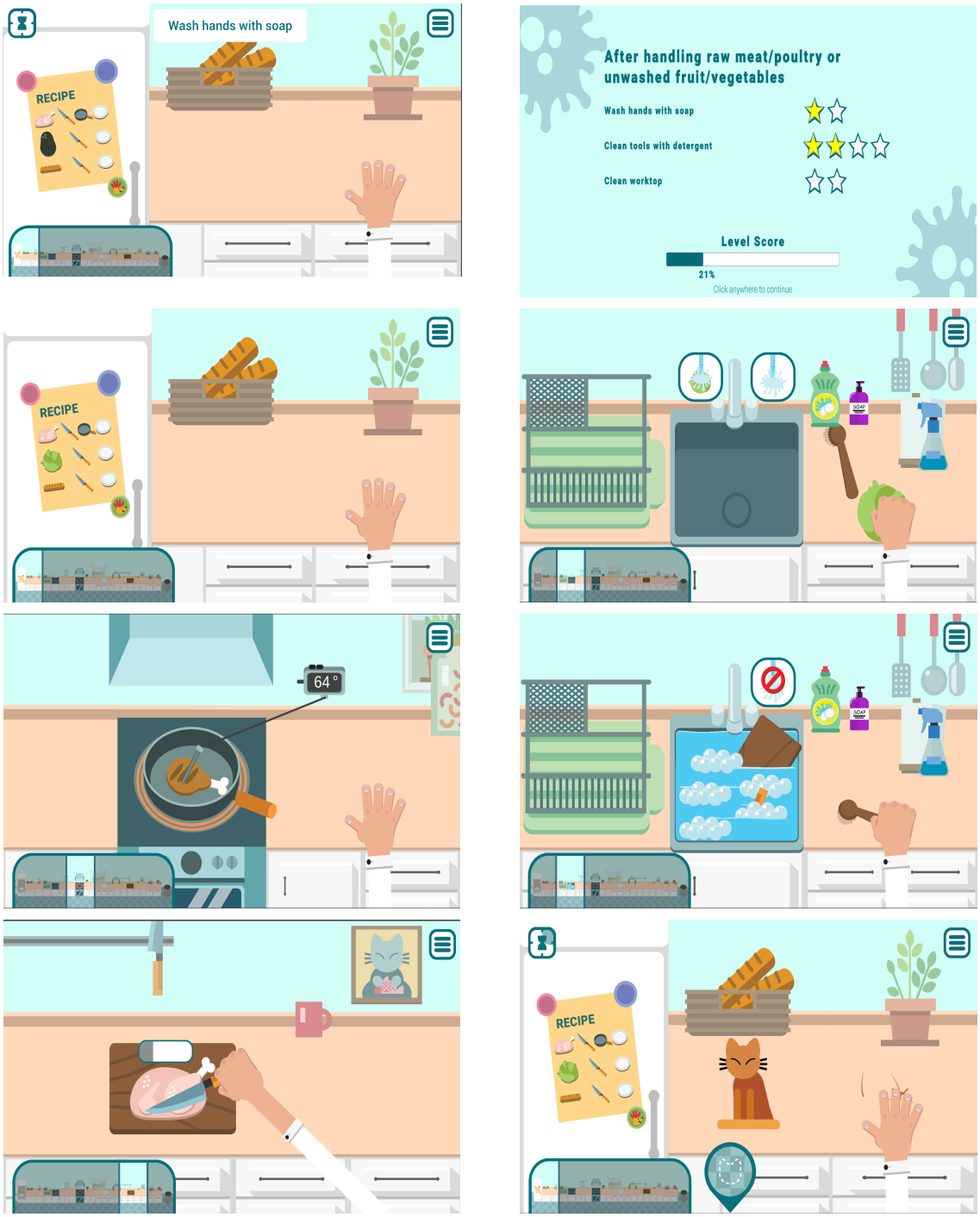


Table 1: Important food safety actions (IFSAs)

1	Wash hands at the start of the recipe
2	Rinse vegetable/fruit
3	Wash hands after raw chicken handled
4	Wash hands after unwashed vegetable/fruit handled
5	Wash knife after raw chicken cut
6	Wash cutting board after raw chicken cut
7	Wash knife after vegetable/fruit cut
8	Wash cutting board after vegetable/fruit cut
9	Clean worktop after chicken cut
10	Clean worktop after vegetable/fruit cut
11	Check with thermometer if chicken reached 74°C
12	Do not rinse raw chicken
13	Throw out dropped bread (only relevant in recipes where bread drops*)

Notes: *In some recipes bread drops to the floor and is visibly dirty when picked up. A correct action is recorded if it is thrown out and replaced by a new bread.

correctly performed. The IFSA score is a percentage score based on how many of the four recipes in a level were completed within the time limit and how many of the IFSAs were correctly performed. In two out of the four recipes, the bread drops. Hence, there are a total of 50 IFSAs and 4 time limits to be complied with in each level.

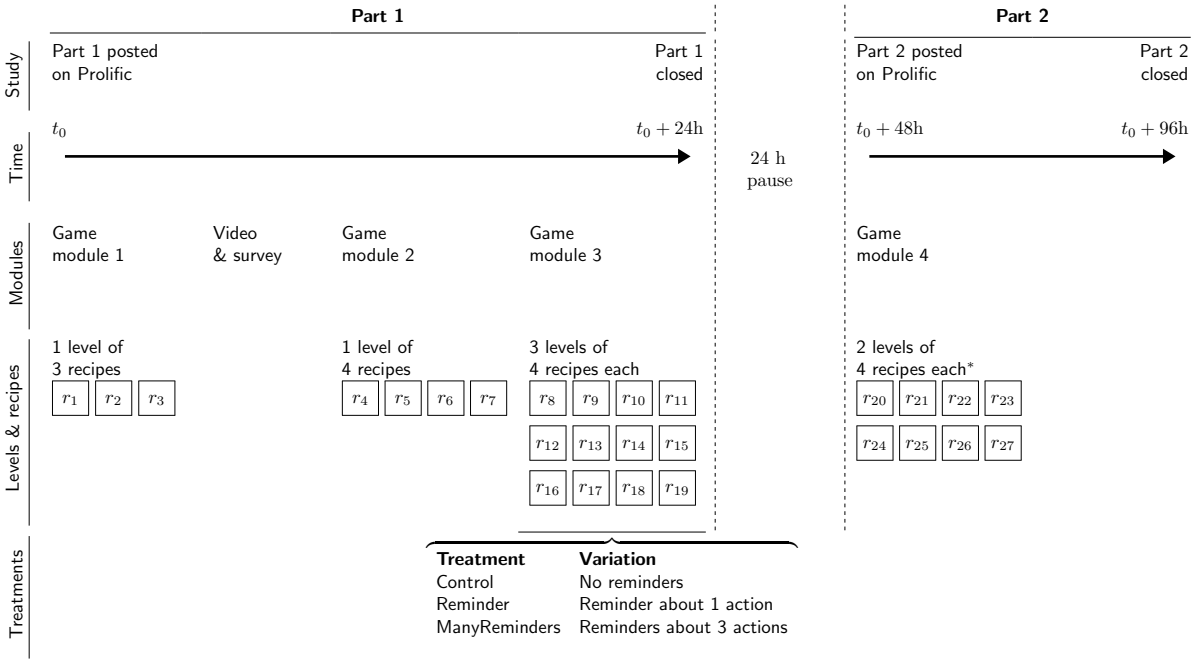
The number of IFSA actions is exogenously fixed by design to avoid that the structure of the feedback depends on the course of action taken during the game, which could be influenced by the treatment. The counter for an IFSA is increased by 1 if an action is performed correctly for the first time in each recipe. That is, repetitions, such as "touch chicken, wash hands – touch chicken, wash hands ..." do not count.

2.2 Timing and treatments

Figure 2 provides an overview of the study. Subjects complete the following sequence over two nonconsecutive dates.

Date 1. Subjects start with **module 1**, where they first see a short video introducing the game mechanics. They then play a practice round of the game without time pressure and subsequently complete two recipes under time pressure . The idea of module 1 is to observe

Figure 2: Overview of the study.



Notes: *The no-feedback study only has four recipes in module 4.

behavior when subjects are not yet influenced by food safety advice or any feedback about their game play.

After this module, subjects complete a **survey**. The survey consists of questions about food safety related knowledge and behavior. Further, we ask about some demographics and preferences. These questions after module 1 provide us with a range of control variables.

After the survey, subjects receive information about safe food handling, addressing all the relevant IFSAs, in the form of a 2-minute **video**. Subjects continue with **module 2**. It starts with a video of a play-through of the game with correct food safety behavior. Subjects then play through four recipes to give us data on their behavior after they have received the food safety advice but before they are incentivized to follow this advice. In module 2, subjects do not get any feedback about their performance.

Module 3 is the first main part of the experiment. Subjects are first told that their payment from now on will depend on their performance. In addition to the fixed payment, there is a bonus of up to GBP 3 based on how well a subject plays the game. Specifically, one of the five game levels that a subject will subsequently play is randomly selected at the end of the study. The bonus is then calculated as GBP 3 $\times$ *score* of the level drawn, where the score assigns a value between 0 and 1 based on which percentage of the four recipes in the level were completed within the time limit and which percentage of the IFSAs were correctly performed. A brief video at the start of module 3 explains how the score is determined and what feedback subjects receive at the end of each level about their performance. Subjects then play through the three levels of module 3. Each level requires completing four recipes.

Module 3 is the stage where the experimental variation of the study happens. Subjects are randomly assigned to one of the three experimental conditions. In the *Control* condition, subjects simply play the game and receive feedback. The two treatments differ from *Control* as follows. In *Reminder*, subjects receive reminders at exogenously fixed points in time about one type of IFSA, namely whenever they are supposed to wash hands (see the top left panel in Figure 1). In total, there are three reminders during a recipe: to wash hands at the start of the recipe and after handling the vegetable and chicken, respectively (see Table 2).

In *ManyReminders*, in addition to the hand-washing reminders, subjects are reminded of two further types of actions: to clean the surfaces after preparing a food item and to check the temperature of the meat. In total, decision makers receive six reminders during a recipe in this treatment (see Table 2).

A reminder appears for few seconds and then disappears again. It only appears when the respective IFSA is relevant for the first time. If for some reason an action sequence is repeated, the reminder is not triggered again. In this way we exogenously fix the number of reminders that the decision maker receives and avoid confounds to our identification strategy. The video that subjects watch at the start of module 3 informs them about the specific actions that they will be reminded of. That is, the reminders are anticipated.

The reminders serve three purposes. First, a reminder on, say, hand-washing should direct attention to this action and thus prompt hand-washing. Second, the reminder may also influence attention to other actions. Third, the reminders make it more likely that the decision maker in fact repeats a particular cue-action sequence, which supports building routines that persist after the reminder is withdrawn.

Date 2. 48 hours after part 1 opened on the Prolific platform, part 2 opens for those subjects who completed part 1. Here subjects complete **module 4**, which is the second main part of the experiment. It is identical across all treatments. Subjects play the game and receive feedback after each of two levels (each level consists of four recipes). That is, the reminders are withdrawn in *Reminder* and *ManyReminders*. After game play, subjects answer a few questions about enjoyment of the game, and possible information seeking regarding food safety during the time of the experiment.

2.3 Procedures

Our study ran on Prolific, with UK residents aged 18-60 years with no dietary restrictions. Further, we required subjects to have a Prolific approval rate of at least 90 percent and to have participated in at least 10 previous studies. The sample was stratified to achieve gender balance. Subjects were informed that they needed earphones or had to activate the speaker,

Table 2: Reminders

Treatment	Description	When shown
Reminder/ ManyReminders	Wash hands with soap	At the start of the recipe
Reminder/ ManyReminders	Wash hands with soap	Vegetable/fruit placed on cutting board
Reminder/ ManyReminders	Wash hands with soap*	Cut meat placed in the pan
ManyReminders	Clean worktop	Cut vegetable/fruit picked up from cutting board
ManyReminders	Clean worktop*	Cut meat placed in the pan
ManyReminders	Check with thermometer that 74 °C reached*	Cut meat placed in the pan

Notes: Reminders are only shown once for each type of action. *Shown together in one pop up message in *ManyReminders*.

and that they could only access the study on a computer.

Subjects earned GBP 8 for day 1 and GBP 6 for day 2 in Fall 2021 (GBP 9 and GBP 3 in Fall 2022) and received a performance dependent bonus of up to GBP 3.

2.4 Amendment: The no feedback study

In the main study, subjects received feedback about their performance in the game. Our pre-registration noted that feedback could be a confounding factor in testing some of our hypotheses, which was partly confirmed when analyzing the data for the feedback study (see Appendix A.2). To address this issue, we designed and pre-registered the *no feedback study* after we analyzed the data of the *feedback study*. The *no feedback study* does not give subjects any feedback on their performance throughout the entire experiment. It is an exact replication of the main study up to the first four recipes of module 3, after which subjects in the *feedback study* for the first time received feedback.

There is a minor modification of the procedure at the start of module 3: The video from the main study that explains the feedback is replaced with an analogous video that explains the IFSA scores on which the payment will be based. Lastly, for budgetary reasons, instead of playing two levels in module 4, subjects play only one level – a minor difference, since the main analysis was originally pre-registered to rely only on the first level of module 4.

Next to being a robustness check for and, for some hypotheses, a replication of the *feedback*

study, the *no feedback study* is also of independent interest: Real world settings in which reminders could be implemented include both cases where individuals frequently receive feedback about their performance (e.g., forgetting an action in a work process typically gets noticed) and where they do not (e.g., omitted safety procedures often go unnoticed until an accident actually happens). Because nudges might have different effects in the two settings, the *no feedback study* makes a contribution in itself.

2.5 Sample

Our final sample consists of the 764 subjects who completed both parts and for whom we have complete data from the game in the feedback study (778 subjects in the no feedback study)[5] Average pay was GBP 15.74 in the feedback study and GBP 14.04 in the no feedback study.[6]

3 Hypotheses

Our hypotheses concern the effects of reminders on immediate behavior (in module 3) and on behavior when reminders no longer are present (in module 4). We compare between control and treatment(s) the number of important food safety actions (IFSAs) performed in the first four recipes of the relevant modules, because subjects receive feedback about their performance after completing these recipes in the feedback study. To test for spillover effects, we distinguish in these analyses between reminded actions (hand-washing in *Reminder*; and hand-washing, cleaning the surface, and checking the temperature in *ManyReminders*) and non-reminded actions (all actions that are not reminded actions in the relevant treatment). Note that the spillover effects are the difference between the effect on all actions and the reminded action(s). To facilitate exposition, we nevertheless report the spillover effects separately in the hypotheses and the tables.

3.1 Immediate effects of reminders

Our first hypothesis is that reminders in module 3 have an overall net positive effect on the total number of correct actions. The positive effect on overall behavior is expected to stem from a direct positive effect on the reminded action. Due the complexity of the environment, we expect that decision makers do not pay attention to all actions. A reminder draws the (limited) attention of a decision maker to the reminded action (see Section 6.2 for further discussion and alternative mechanisms). Consequently, we expect that reminders help the

[5] 868 subjects completed the feedback study (823 subjects in the no feedback study). 972 (842) subjects only completed part 1. Because of server connection issues, some or all game data was not saved for 104 (74) subjects.

[6] Prolific raised the minimum payment for participants between our two experiments from GBP 5/h to GBP 6/h in April 2022. The no feedback study however was somewhat shorter.

decision maker to perform the correct actions about which they are reminded in *Reminder* more often compared to what the decision makers in *Control* do. This leads to Hypothesis 1 (i) below.

The direction of the spillover effects in *Reminder* is less clear. On the one hand, reminders may have negative spillover effects on other actions (crowding out). First, reminders may induce attention overload (e.g., Cattaneo et al., 2021) or be a distraction. Second, performing the reminded actions may induce subjects to perform fewer of the other actions because of (mental or effort) cost spillovers, or because of licensing effects (Merritt et al., 2010). On the other hand, reminders may have positive spillover effects. First, because subjects know that they will be reminded of hand-washing, this potentially frees up attention resources that decision makers can then direct towards the non-reminded actions. The psychology literature refers to this as cognitive or intention offloading (Risko and Gilbert, 2016; Gilbert et al., 2023). Heidhues et al. (2021) propose a similar positive attention spillover effect in a theoretical model of consumer regulation. Second, the similarities of some of the food safety actions (such as washing hands and washing kitchen utensils) might help decision makers who see a reminder to wash hands, to also recall to wash utensils (for overviews of theories of memory and recall, and the role of similarity see Kahana, 2020; Malmendier and Wachter, 2021). Whether negative or positive spillover effects dominate is therefore an open empirical question. That is why Hypothesis 1 (ii) is not directional.

For overall performance, however, we expect a positive effect of reminders. It seems unlikely that decision makers are at their capacity limit in terms of effort or attention (otherwise providing one action more would force the individual to provide one action less of another type). Therefore, even if spillover effects potentially are negative, we do not expect them to dominate the positive direct effect of reminders.

Hypothesis 1 *In the first four recipes of module 3, the decision makers perform more correct actions in Reminder than in Control. The overall effect is decomposed into:*

(i) Direct effect of the reminders: The decision makers perform more often the correct actions about which they are reminded in Reminder than in Control.

(ii) Indirect effects of the reminders: The decision makers perform more or less often the correct actions about which they receive no reminders in Reminder than in Control.

Our second hypothesis concerns the scalability of reminders to multiple actions for which we compare treatment *ManyReminders* with *Reminder*. The reminders for additional actions that we give in *ManyReminders* in comparison to *Reminder* are expected to have a positive effect on these targeted actions, unless they lead to (information) overload so that decision makers start to ignore them. Further, by similar arguments as above, both positive as well as negative spillover effects could arise. The negative spillover effects could be more pronounced with *ManyReminders* than with *Reminder* because more reminders are more likely to increase

cognitive load and crowd-out attention. It is an empirical questions whether these negative spillovers (if they arise) are more pronounced in *ManyReminders* than in *Reminder* and, if so, whether they dominate the positive direct effect. That is why Hypothesis 2 is not directional.[7]

Hypothesis 2 *In the first four recipes of module 3, the decision makers perform more or less often the correct actions in ManyReminders than in Reminder.*

If we find a positive effect, then as a secondary hypothesis, we will look separately at the direct and spillover effects of *ManyReminders* vs. *Reminder*.

3.2 Intertemporal effects of reminders

Our third hypothesis concerns the intertemporal spillover effects of being treated with reminders in module 3 on actions in module 4, where reminders no longer are provided. Because we expect a positive effect of reminders on the actions about which decision makers are reminded in module 3 (see Hypothesis 1i), we expect that decision makers in *Reminder* will be more likely to continue to wash hands in module 4 than decision makers in *Control*. Repetitions of a cue-action series help build routines that lead to persistent behavior; and such repetitions are more likely to take place if the decision maker receives reminders that prompt the action.[8] If the hypothesis of a persistent positive direct effect is not rejected, we will test whether reminders also have a net positive impact when considering all actions. The spillover effect is again the difference between the effect on all actions and the reminded actions. In parallel to Hypothesis 1, we hence state:[9]

Hypothesis 3 *In the first four recipes of module 4, the decision makers perform more correct actions in Reminder than in Control. The overall effect is decomposed into:*

(i) Direct effects of the reminders: The decision makers perform more often the correct actions about which they are reminded in Reminder than in Control.

(ii) Indirect effects of the reminders: The decision makers perform more or less often the correct actions about which they receive no reminders in Reminder than in Control.

[7] The hypotheses follow the pre-analyis plan. Had we written up this hypothesis ex post, we would have written it up in the same way as Hypothesis 1.

[8] We speak here of routines or persistent behaviors because the literature on habit formation typically considers a longer time frame. Yet, the ideas are related to habit formation. Gardner (2015, p.280) defines a habit as "a process by which a stimulus automatically generates an impulse towards action, based on learned stimulus-response associations". The psychology literature emphasizes that in order to create habits it is important to repeatedly apply an action in response to a cue and to receive rewards for taking the action (e.g., Wood and Neal, 2007, 2009). Hussam et al. (2022) points out that hand-washing is an ideal candidate for habit formation because of its repeated nature and because it is associated with clear cues. The same applies in the game. We reinforce the cues by sending reminders that appear at pre-determined events where a food safety action should be taken. The reward comes in form of a higher score in the feedback stage.

[9] Hypothesis 3 is labelled as Hypotheses 3 and 4 in the pre-analysis plan.

Our final hypothesis addresses the question whether the effects of reminders scale up even after the nudge no longer is present.

Hypothesis 4 *In the first four recipes of module 4, the decision makers perform more or less often (with the same direction as in the empirical result for Hypothesis 2) the correct actions in ManyReminders than in Reminder.*

3.3 Secondary hypotheses

For all hypotheses comparing *Reminder* vs. *Control*, we test the analogous hypotheses for *ManyReminders* vs. *Control*. In addition, we test the effects on the reminded and non-reminded actions in the comparisons between *ManyReminders* and *Reminder*. The secondary hypotheses, and the additional analyses that we perform to provide further insights into possible mechanisms underlying the main results, were not part of the ex ante power analysis that served as the basis for the pre-analysis plan.

4 Empirical strategy

We estimate the treatment effects in our experiment using the regression analysis of covariance (ANCOVA) model. This model conditions on pre-treatment measurements to increase the precision of the treatment estimator and is more efficient than the alternative of a difference-in-differences estimator (McKenzie, 2012; Burlig et al., 2020; Ek, 2020). Specifically, we estimate the following equation via a least squares regression:

$$Y_{i,t} = \alpha_t + \boldsymbol{\beta}\,\mathbf{X}_i + \gamma\,T_{i,t} + \theta\,\bar{Y}_{i,Pre} + \epsilon_{i,t}, \tag{1}$$

where i refers to the individual and the post-treatment date t to the recipe number, $Y_{i,t}$ is the outcome variable of individual i at date t, α_t are recipe fixed effects, $\mathbf{X}_i$ is a set of predetermined control variables, $T_{i,t} \in \{0, 1\}$ is the treatment status of i at date t, and $\bar{Y}_{i,Pre}$ is the mean for individual i over the pre−treatment measures in modules 1 and 2.

Our main outcome variable is the average IFSA score of a recipe. This is the sum of the binary scores for the 12 IFSAs (13 IFSAs in recipes where bread drops) in Table 1. As outlined above, we further distinguish between actions about which the decision maker is reminded in module 3 and actions about which the decision maker is not reminded.

We report three specifications: (i) without covariate vector $\mathbf{X}_i$, (ii) with the set of basic controls, and (iii) with the set of extended controls. We describe the control variables in Appendix A.1.[10]

[10]Because of a software error, Qualtrics did not record for a few subjects (Feedback study: 51 subjects – 20 control, 11 Reminder, 20 ManyReminders. No feedback study: 50 subjects – 16 control, 18 Reminder, 16 ManyReminders) the survey question on prior use of a food thermometer at home that is needed to construct

Given that the treatments are expected to affect the score that the subjects get feedback on in modules 3 and 4 in the *feedback study*, the score could also influence motivation and hence performance in subsequent play. For that reason, our hypotheses are stated for the first four recipes of module 3, where subjects have not yet received any feedback in the *feedback study*. In addition, the *no feedback study* is a replication of the *feedback study* for the analyses based on data from the first four recipes of module 3. For the results related to module 4 (Hypothesis 3-4), the *no feedback study* serves as a check that the results of the *feedback study* are not confounded by feedback on game performance that subjects received during module 3.

5 Results

5.1 Descriptive statistics

Table 3 provides descriptive statistics for the average IFSA score per recipe for all actions, the reminded actions, and the non-reminded actions in the different modules. Figures A.1-A.10 show the evolution of the average IFSA scores in the different treatments over the different recipes. The samples look reasonably balanced according to Tables A.5-A.6. As expected with multiple t-tests, a few variables show statistically significant differences across treatments. Our empirical approach controls for such imbalances. In any case, we will see that our results are robust across specifications with and without controls.

5.2 Immediate effects of reminders (Hypotheses 1 and 2)

***Reminder* vs *Control* (Hypothesis 1).** Columns 1 to 3 of Table 4 show that subjects in *Reminder* more often take the correct food safety actions than subjects in *Control* – both in the *feedback study* (Panel A) and the *no feedback study* (Panel B). Columns 4 to 6 reveal that the positive treatment effect stems from the actions that subjects are reminded about in *Reminder*. In contrast, Columns 7 to 9 show a negative treatment effect on the non-reminded actions, but it is only significant in the *no feedback study*.

The patterns are similar for *ManyReminders* vs. *Control* (Secondary Hypothesis 1, Table A.7). There is a positive treatment effect of reminders on overall performance, driven by a positive direct effect on the reminded actions that is strong enough to compensate for the negative spillover effect on the non-reminded actions. All effects are robustly significant across the *feedback* and *no feedback* studies.

Summarizing the findings, reminders induce subjects to take the targeted actions, but crowd-out non-targeted actions. Crowding-out is not a phenomenon that appears because of aggre-

the pre-registered control variable *PriorBehaviors*. We report results using imputed values based on the other observables, but results are unaffected when using only the sample for whom the question was recorded or when dropping the *PriorBehaviors* variable.

Table 3: Descriptive statistics for the number of correct actions (IFSA score)

IFSA score											
Total (all actions)				Reminded actions[b]		Non-reminded actions[c]		Reminded actions in ManyReminders[d]		Never-reminded actions[e]	
Module[a]											
1	2	3	4	3	4	3	4	3	4	3	4
Control (N=255), Feedback study											
3.73	6.47	8.16	9.13	1.65	6.51	1.94	7.19	3.18	3.77	4.98	5.36
1.77	2.79	2.75	2.49	0.87	2.17	0.81	1.94	1.48	1.35	1.56	1.39
Control (N=252), No feedback study											
3.85	6.76	8.48	8.41	1.67	6.81	1.72	6.69	3.35	3.38	5.13	5.04
1.76	2.58	2.63	2.52	0.93	1.99	0.85	1.93	1.45	1.41	1.43	1.37
Reminder (N=255), Feedback study											
3.90	6.79	9.15	9.25	2.63	6.52	2.16	7.09	4.15	4.02	4.99	5.23
1.68	2.82	2.44	2.49	0.70	2.12	0.76	2.04	1.18	1.34	1.53	1.39
Reminder (N=271), No feedback study											
3.97	6.96	9.27	8.51	2.73	6.54	2.07	6.44	4.34	3.63	4.93	4.88
1.81	2.77	2.35	2.65	0.56	2.10	0.82	2.13	1.12	1.41	1.47	1.51
ManyReminders (N=254), Feedback study											
3.91	6.80	9.73	9.22	2.54	7.19	2.04	7.18	4.90	4.12	4.83	5.10
1.80	2.74	2.33	2.67	0.67	1.98	0.84	2.10	1.16	1.44	1.61	1.56
ManyReminders (N=255), No feedback study											
3.79	6.44	9.31	8.30	2.53	6.78	1.89	6.41	4.79	3.78	4.52	4.52
1.80	2.87	2.49	2.88	0.70	2.07	0.92	2.18	1.28	1.60	1.61	1.65

Notes: Mean (std.dev.) of the IFSA score. Overall number of participants N=764 (Feedback study) and N=778 (No feedback study). [a]For the first four recipes of modules 3 and 4, respectively. That is, before any feedback on the IFSA score is given in the feedback study. [b]The category of actions reminded in *Reminder*: Hand-washing. [c]All actions except hand-washing. [d]The categories of actions reminded in *ManyReminders*: Hand-washing, surface cleaning, and checking the temperature of the meat. [e]All actions except surface cleaning, and checking the temperature of the meat.

gation, but shows across the different actions: When looking at actions separately, we do not observe any positive signs for treatment effects on non-reminded actions (see Tables A.8 and A.9). For example, one might have expected that a reminder on hand-washing could have a positive spillover effect on actions which are associated with hand-washing, such as washing the cutting board and knife. But this is not the case.

***ManyReminders* vs *Reminder* (Hypothesis 2).** We next ask whether one can scale-up the positive effect of a reminder by giving reminders for more types of actions. Table 5 shows that subjects indeed more often perform the correct actions in *ManyReminders* than in *Reminder*. Table A.13 unpacks this overall effect into direct and spillover effects (Secondary Hypothesis 2). Columns 4 to 6 show that subjects perform better because they are reminded about more actions in *ManyReminders* (hand-washing, cleaning the worktop and checking the meat temperature) than in *Reminder* (hand-washing only). Yet, Columns 1 to 3 reveal a negative spillover on the effectiveness of the hand-washing reminder – the action reminded in both treatments: Subjects in *ManyReminders* tend to do worse in hand-washing than subjects in *Reminder*. Below, we investigate this further. In addition, the negative spillover effects of reminders on non-targeted actions are more severe when subjects receive more reminders. Columns 7 to 9 show that that subjects in *ManyReminders* do worse on the actions that are not reminded in any treatment than subjects in *Reminder*.

5.3 Intertemporal effects of reminders (Hypotheses 3 & 4)

We now consider the intertemporal spillover effects that reminders have on behavior once the nudges are withdrawn in module 4.

***Reminder* vs *Control* (Hypothesis 3).** Table 6 shows that reminders are – at least to some extent – successful in creating persistent behavior. Columns 4 to 6 show that subjects in *Reminder* – who were treated with reminders in module 3 – still wash their hands more often than subjects in *Control* do, even though they no longer are reminded to do so. However the effect of treatment with reminders partially fades out compared to the impact of an active reminder: The coefficient of the treatment dummy for *Reminder* in module 4 is less than one third of the respective coefficient for *Reminder* in module 3.[11]

Remarkably, there is no such fade-out for the negative spillover effect on the non-reminded actions: There is no significant difference in the coefficients for the treatment effect of *Reminder* vs. *Control* on the non-reminded actions in module 4 compared to module 3.[12] This finding suggests that once subjects forget about performing an action because a reminder induced them

[11] The respective coefficients in Columns 4 to 6 in Tables 4 and 6 are significantly different according to a Wald chi-square test ($p < .0001$ for all specifications in both the *feedback* and *no feedback* studies).

[12] According to a Wald chi-square test comparing the respective coefficients in Columns 7 to 9 of Tables 4 and 6 ($p = 0.4460/p = 0.9316$ for the *feedback/no feedback* study with the full set of controls).

Table 4: Hypothesis 1

Panel A: Feedback study (510 individuals; N=2040)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.74***	0.77***	0.76***	0.93***	0.92***	0.91***	-0.19	-0.15	-0.16
vs. Control	(0.15)	(0.14)	(0.14)	(0.06)	(0.06)	(0.06)	(0.12)	(0.12)	(0.12)
Adj. R2	0.55	0.60	0.62	0.35	0.37	0.39	0.53	0.59	0.60
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.16	8.16	8.16	1.65	1.65	1.65	6.51	6.51	6.51
Std.dev.	2.91	2.91	2.91	1.02	1.02	1.02	2.30	2.30	2.30
Panel B: No feedback study (523 individuals; N=2092)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.64***	0.60***	0.63***	1.03***	1.01***	1.01***	-0.39***	-0.41***	-0.38***
vs. Control	(0.14)	(0.13)	(0.13)	(0.06)	(0.06)	(0.06)	(0.12)	(0.11)	(0.11)
Adj. R2	0.52	0.60	0.61	0.40	0.43	0.44	0.49	0.56	0.57
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.48	8.48	8.48	1.67	1.67	1.67	6.81	6.81	6.81
Std.dev.	2.78	2.78	2.78	1.05	1.05	1.05	2.13	2.13	2.13

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except hand-washing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table 5: Hypothesis 2

Panel A: Feedback study (509 individuals; N=2036)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	0.57***	0.57***	0.58***
	(0.14)	(0.13)	(0.13)
Adj. R2	0.50	0.56	0.57
Controls	No	Basic	Ext
Mean (Reminder)	9.15	9.15	9.15
Std.dev.	2.59	2.59	2.59

Panel B: No feedback study (526 individuals; N=2104)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	0.36***	0.43***	0.42***
	(0.13)	(0.12)	(0.12)
Adj. R2	0.52	0.59	0.59
Controls	No	Basic	Ext
Mean (Reminder)	9.27	9.27	9.27
Std.dev.	2.50	2.50	2.50

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: Overall IFSA score. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

to take another action, they also do not remember to perform it later on, when reminders are withdrawn. That is, fading-out of the effect of reminders does not imply crowding-in of the non-reminded actions.

Columns 1 to 3 in Table 6 consider the impact on the overall performance in module 4 of having been treated with reminders in module 3. There is no significant difference between *Reminder* and *Control*. The persistent negative spillover effect on the non-reminded actions completely offsets the diminished positive effect on the reminded actions.

***ManyReminders* vs *Reminder* (Hypothesis 4).** Table 7 considers how scaling up the number of reminded categories in module 3 affects overall performance after reminders are withdrawn in module 4. We observe no significant difference between *ManyReminders* and *Reminder*. Thus, the scaling effects on overall performance are not persistent. , Table A.14 unpacks this overall effect into direct and spillover effects (Secondary Hypothesis 4). We again see partial fade-out of the effects on the reminded actions and persistence of the spillover effects on the non-reminded actions, but they are not robustly significant across the *feedback* and *no feedback* studies, though the signs are robust. We further explore the role of feedback in Section 6.1.1.

5.4 Robustness

We summarize here multiple analyses which show that our results are robust. In the *feedback study*, from module 3 onward subjects receive feedback about their performance in the game after each set of four recipes. This is why we only used the performance data for the first four recipes of each module. Tables A.15-A.20 repeat our main analyses using all recipes within a module.[13] Another robustness check uses an alternative outcome measure (see Appendix A.4). Finally, we check for a possible confound if treatment with reminders induces subjects to access food safety information more often than subjects in *Control*. In the post survey, we ask subjects how often they accessed food safety information (outside of the study information) during the period of the study. Further, we observe how often subjects clicked to reopen any videos and how much time they spent on the videos. There is no significant difference across conditions in all these variables (see Appendix A.5).

6 Additional analyses

In this section we present some exploratory analyses that highlight certain mechanisms, provide more insights into our results, and allow us to discuss alternative mechanisms.

[13] In module 3, there are 12 recipes and in module 4 there are eight (four) recipes in the feedback study (no feedback study). The last four recipes were dropped in the no feedback study for cost reasons.

Table 6: Hypothesis 3

Panel A: Feedback study (510 individuals; N=2040)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	-0.08	-0.07	-0.09	0.17***	0.16***	0.17***	-0.25*	-0.24*	-0.25*
vs. Control	(0.17)	(0.16)	(0.16)	(0.06)	(0.06)	(0.06)	(0.14)	(0.13)	(0.13)
Adj. R2	0.38	0.43	0.44	0.17	0.19	0.20	0.36	0.41	0.43
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	9.13	9.13	9.13	1.94	1.94	1.94	7.19	7.19	7.19
Std.dev.	2.66	2.66	2.66	0.96	0.96	0.96	2.08	2.08	2.08
Panel B: No feedback study (523 individuals; N=2092)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	-0.06	-0.10	-0.06	0.31***	0.30***	0.30***	-0.37***	-0.39***	-0.37***
vs. Control	(0.16)	(0.15)	(0.16)	(0.06)	(0.06)	(0.06)	(0.13)	(0.12)	(0.13)
Adj. R2	0.45	0.51	0.52	0.23	0.27	0.28	0.43	0.49	0.49
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.41	8.41	8.41	1.72	1.72	1.72	6.69	6.69	6.69
Std.dev.	2.66	2.66	2.66	0.96	0.96	0.96	2.05	2.05	2.05

Notes: Data are from the first four recipes of module 4. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except hand-washing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table 7: Hypothesis 4

Panel A: Feedback study (509 individuals; N=2036)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	-0.04	-0.02	0.02
	(0.17)	(0.16)	(0.16)
Adj. R2	0.41	0.45	0.46
Controls	No	Basic	Ext
Mean (Reminder)	9.25	9.25	9.25
Std.dev.	2.68	2.68	2.68

Panel B: No feedback study (526 individuals; N=2104)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	0.14	0.18	0.17
	(0.16)	(0.16)	(0.16)
Adj. R2	0.48	0.53	0.54
Controls	No	Basic	Ext
Mean (Reminder)	8.51	8.51	8.51
Std.dev.	2.78	2.78	2.78

Notes: Data are from the first four recipes of module 4. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: Overall IFSA score. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

6.1 Additional analyses related to the hypotheses

6.1.1 Hypotheses 3-4: Why does the positive effect of reminders fade out?

Our analyses below provide some indication for why the effect of reminders partially fades out: some of their impact appears to be driven by a mechanical effect of bringing an action to the top of the mind of the decision maker in the moment that the reminder appears. In addition, receiving too many of reminders diminishes their effectiveness.

The effect of feedback. In the feedback study, one potential explanation for why treatment differences in overall performance vanish over time might simply be that subjects in *Control* become better because they learn from the feedback about their game performance. To examine this possibility, we consider the evolution of the treatment effects over the course of module 3. In all comparisons feedback increases performance over time (see Tables A.30-A.32).[14] Further, the significant negative coefficient on the interaction between feedback and treatment indicates that the gap between *Control* and the other treatments narrows over time. Appendix A.2 provides further analysis of these patterns. Overall, feedback and reminders seem to partially substitute each other. Subjects appear to learn from the feedback that they receive and the feedback may also help subjects pay attention to the correct actions. Hence, feedback contributes to the fading-out of the effect of reminders. Yet, feedback cannot be the main driver – otherwise we would not observe the same main results in the *no feedback study*. We therefore turn to different explanations next.

The mechanical effect of reminders. In the absence of feedback, performance in module 3 does not evolve differently over time in *Reminder* and *Control*: The treatment dummy for the comparison between *Reminder* and *Control* is not significant in Columns 1 and 2 of Table A.30 that analyses changes in performance over time. This, together with the fact that reminders immediately increase overall performance and performance on the reminded actions (see Section 5.2) and the results from Section 5.3, hints at a fairly mechanical effect of a reminder bringing an action to the top of the mind of the decision maker in the moment that the reminder appears. An explanation for why the treatment effect of reminders fades out after reminders are withdrawn could be that because subjects do not see a reminder, they start forgetting about performing the action and fall back into their old (real world) routines. Yet, the fade-out of reminders does not seem to be entirely driven by the mechanical effect above as we now discuss.

[14] Specifically, we regress the difference in the average IFSA scores per recipe between the last four recipes in module 3 (the last level before reminders are withdrawn) and the first four recipes of module 3 (the level before feedback is given for the first time in the feedback study) on a treatment dummy, a dummy for whether the subject participated in the feedback study or not, and the interaction between the feedback and treatment dummies.

Decreasing responsiveness. When reminders are given for many categories, they become less effective over time. Specifically, the comparisons between *ManyReminders* and *Control* in Table A.31 reveal a negative and significant treatment dummy in regressions where the dependent variable is the difference over time in the overall IFSA score (Column 1) or in the score for the reminded actions (Column 2). Being exposed to many reminders may lead to decreasing responsiveness because of attention overload or simply because reminders become an annoyance, both of which would make the reminders less effective over time and less helpful in training persistent behavior. This explanation is consistent with Archsmith et al. (2021) and Embi and Leonard (2012), who provide evidence that is suggestive for the interpretation that attention depletes over time in response to previous actions or alerts.[15]

New routines are difficult to establish. Some actions may be easier to forget than others. For example, after reminders are withdrawn in module 4, we observe a drop in hand-washing and surface cleaning after handling vegetables and meat compared to module 3 (see Appendix A.6). The drop tends to be larger for these behaviors after handling vegetables compared to after handling meat, though the effect is not consistent across all comparisons. Food safety is obviously more important after handling raw meat and, therefore, it is likely that already prior to the study subjects had a routine of hand-washing and cleaning after handling meat. While these actions are also recommended after handling unwashed vegetables,[16] individuals may have had a routine of *not* doing anything after handling vegetables. So the reminders of hand-washing and cleaning after handling meat are likely to reinforce something that subjects already knew they should do. In contrast, the reminders of hand-washing and cleaning after handling vegetables are likely to go against old routines and, therefore, possibly less successful in helping subjects establish new routines.[17]

6.1.2 Hypothesis 2: Scaling

The positive treatment effect of *ManyReminders* vs. *Reminder* on performance in Section 5.2 suggests that one can scale-up the immediate effects of reminders to more reminded categories, because the reminders crowd-out neither each other *one-to-one*, nor other behaviors. Yet, our data also suggest that reminders have decreasing returns to scale. First, the immediate effect of reminders on the respective IFSA score per reminded action is smaller in *ManyReminders* than in *Reminder*.[18] Second, the negative coefficient on the treatment dummy for *ManyReminders*

[15] A competing story as to why we would observe such a relationship is that effort or mental costs increase over time, for example due to fatigue. This is however difficult to reconcile with the fact that there is no significant decay in performance in *Control* over time – unless there are discontinuities in the cost function.

[16] See, for example, https://www.nhs.uk/live-well/eat-well/how-to-wash-fruit-and-vegetables/ (last accessed April 2021).

[17] See Betsch et al. (2015) and their discussion of the literature on the difficulty of breaking with old routines.

[18] Considering that there are six reminded actions per recipe in *ManyReminders*, the treatment dummy in module 3 for *ManyReminders* vs. *Control* (1.67 in the feedback study and 1.55 in the no feedback study; see

vs. *Reminder* in Table A.13 suggests that hand-washing reminders become less effective when more categories are reminded (see Columns 1 to 3; the effect however is not robustly significant).

This section looks into possible sources of such decreasing returns to scale. Does a single reminder in *ManyReminders* become less effective because more action categories are reminded and the decision maker receives more reminders in total? Does it become more difficult to effectively handle multiple reminders at the same time? To investigate these mechanisms, we use the fact that in *ManyReminders* subjects receive six reminders distributed across four pop-up messages as follows: One message in the beginning (wash hands), one when placing the vegetable on the cutting board (wash hands), one when placing the vegetable on the plate (clean surface), and three when placing the meat in the pan (wash hands, clean surface, check temperature).

Table A.33 considers performance for these different types of reminded actions in module 3. Columns 1 to 6 list those occurrences where only one hand-washing reminder appears and reveal no significant differences in hand-washing between *ManyReminders* and *Reminder*. In contrast, after handling meat, subjects in *ManyReminders* get reminded about hand-washing at the same time as they get reminders for two other types of actions. Columns 7 to 9 reveal that subjects here perform worse compared to subjects in *Reminder*, who are only reminded of hand-washing: The treatment dummy *ManyReminders* vs. *Reminder* is negative and robustly significant in the two studies. Moreover, the difference in treatment effects between *ManyReminders* vs. *Reminder* for hand-washing after handling meat compared to hand-washing at the start of a recipe or hand-washing after handling vegetables, respectively, is robustly significant according to a Wald chi-square test (all comparisons $p < 0.0001$).

Table A.34 reports a similar analysis of cleaning the surface after handling vegetables (reminder message with one action in *ManyReminders*) compared to cleaning the surface after handling meat (reminder message with multiple types of actions in *ManyReminders*). *ManyReminders* has a larger treatment effect in the former situation than in the latter, according to a Wald chi-square test (all comparisons $p < 0.0001$). A caveat is that the outcome measure in the baseline treatment *Reminder* is lower after handling vegetables than after meat (feedback study 0.17 vs. 0.58; no feedback study 0.21 vs. 0.64). But the difference in treatment effects is still significant when accounting for the difference in the baselines (all comparisons $p < 0.0001$).

Column 6 in Table A.7) is proportionately smaller than that for *Reminders* vs. *Control*, which has three reminded actions per recipe (0.91 in the feedback study and 1.01 in the no feedback study; see Column 6 in Table 4). Because the three additional reminders in *ManyReminders* are on different categories than hand-washing, this comparison is suggestive only.

6.2 Alternative mechanisms

We pursued the idea that subjects fail to engage in some behaviors because they pay limited attention to performing the correct actions. A reminder directs attention to the reminded action and, in doing so, prompts the decision maker to perform the action. In this section, we consider possible alternative mechanisms for why subjects may fail to take correct actions or why reminders may improve performance.

6.2.1 Reminders as a signal and experimenter demand

A potential alternative mechanism for why reminders 'work' is that they signal to the individual that a specific action is of particular importance. The fact that all actions come with the same monetary reward should speak against this interpretation. But subjects may miss this information. Under this alternative mechanism, we would expect reminders to be *more* effective for subjects who pay little attention to the study information. As proxies for attention to the study instructions we use either the total time that an individual spends watching the information videos or the number of times that these videos were opened; and we interact the respective variable with the treatment dummy *Reminder* vs. *Control*.

We find no evidence for the alternative mechanism, or a knowledge channel. First, if more attention to the study instructions lead to better knowledge about what food safety actions should be performed, the overall score in *Control* should increase. Yet, the main effect of the respective proxy for attention is either non-significant or negative.[19] Second, if a lack of attention to the study instructions lead subjects to falsely attribute greater importance to the reminded actions, reminders should be less effective for subjects who spent more time with the information material. Yet, the interaction term is non-significant or positive.[20]

Of course, subjects might well understand the monetary incentives, but nevertheless follow the reminder because of an experimenter demand effect. The evidence in De Quidt et al. (2018) suggests that the demand effects in typical real effort experiments are likely to be small. Nevertheless, to account for a possible experimenter demand effect, we exclude those subjects from the analysis who comply with all reminders in (i) all recipes or (ii) more than half of the recipes of the relevant level. We find no evidence for experimenter demand as a mechanism because our results are robust in terms of signs and statistical significance, with one exception (see Tables A.35 and A.36): Once we exclude those subjects who complied with the reminders in more than half of the recipes, the overall positive effect of *Reminder* vs. *Control* disappears (see Column 2 in Table A.35). The reason is that the positive direct effect of reminders on the reminded action mechanically becomes smaller – because we excluded the sizeable group of

[19] Feedback study/no feedback study $\beta_{\text{duration}} = -0.002/-0.001$, $p = 0.352/0.512$ and $\beta_{\text{clicks}} = -0.08/-0.312$, $p = 0.038/0.004$).

[20] Feedback study/no feedback study $\beta_{\text{duration}\times\text{treatment}} = 0.002/-0.0002$, $p = 0.496/0.907$ and $\beta_{\text{clicks}\times\text{treatment}} = 0.080/0.292$, $p = 0.159/0.038$).

those who complied with the reminder in more than half of the recipes – and now the negative spillover effect of reminders on the non-reminded actions fully offsets it.

6.2.2 Spillovers arising from substitution

An alternative mechanism for why reminders would lead to spillover effects on non-targeted actions could be some form of substitutability of actions in the utility function. One prominent form of such spillovers is licensing (Merritt et al., 2010; Tiefenbeck et al., 2013). Decision makers may have a reference point for how many food safety actions to perform and keep track of them in a mental account (Thaler, 1985) for the recipe, the entire level, or even for several levels. If decision makers do one more of the reminded actions, this 'allows' them to do one less of the non-reminded actions.[21] The fact that we observe a positive treatment effect of reminders on overall performance in module 3 means that, on average, crowding out is not one-to-one – speaking against the view that *all* subjects engage in licensing.

If *some* decision maker engaged in licensing, we would expect crowding-out especially for actions that are not incentivized, such as removing the cat from the kitchen counter top. Yet, we observe no evidence of this. Being treated with reminders has no impact on the number of times that the cat is removed (see the third-last column in Table A.8).

We find no evidence of licensing over time either. Performance does not decay over time in module 3 in *Reminder* and *Control* (cf. Section 6.1.1), speaking against subjects compensating for a 'good' score in one level by slacking in another level.

6.2.3 Ability, knowledge, and preferences

Competing explanations of why a decision maker may fail to take correct actions are (i) lacking ability, (ii) limited knowledge, or (iii) preferences. In our regressions, we control for how subjects play the game in modules 1 and 2, which should reduce and potentially eliminate the possible influence of prior knowledge of food safety and preferences. We address potential indirect effects of these factors on treatment effects in Section 6.2.4.

In any case, our study design limits the role of these factors. First, regarding (i), by the time subjects reach module 3, they should know how to play the game because they have completed seven recipes and watched two videos on how to play in modules 1 and 2. Second, regarding (ii), by the time subjects reach module 3, they should know about the required food safety actions from the video in module 2 and the video-play-through with correct food safety actions at the start of module 3 that provided subjects with knowledge about what actions

[21] We see evidence that subjects apply some of their real world habits in the game, such as using a safer preparation order than the order in the recipe pinned to the fridge (see Appendix A.7). In the context of real world food preparation, licensing or mental accounting effects are unlikely to arise because food safety actions cannot be substituted for each other. For a dish to be safe to eat, all food safety actions have to be followed.

to perform.[22] Third, regarding (iii), we induce preferences for taking the correct actions in modules 3 and 4 by providing incentives tied to the number of correct actions performed; and these incentives appear to work.[23]

6.2.4 Treatment interactions with knowledge or gaming experience

A potential alternative mechanism is that reminders impact performance indirectly by affecting knowledge about food safety actions or how subjects learn when playing the game. When testing our hypotheses, we already control for knowledge through the survey-based *Prior-Knowledge* variable and for experience with the game prior to the treatment with reminders through the IFSA score in modules 1 and 2.

To further explore the potential impact of reminders on knowledge and experience, we perform subgroup analyses. If reminders induced knowledge or helped with building gaming experience, then certain subgroups might respond more or less to reminders. However, we observe no differences in treatment effects for subgroups based on overconfidence[24], experience with playing computer games, or knowledge about food safety. Specifically, we perform the same regressions as for Hypothesis 1 but include either the variable overconfidence, a dummy for people who are more knowledgeable than the median, or a dummy for people who play more frequently computer games than the median, and the interaction between the respective variable and the treatment dummy for *Reminder* vs. *Control*. In all regressions, the interaction terms are non-significant.[25] Overall, this suggests that the effect of reminders is not moderated by knowledge or experience with game play.

6.3 Further exploratory analyses

In this section, we report on further exploratory analyses of the mechanisms driving our results.

[22] In Koch et al. (2022), we provide evidence from an experimental study that the information video on its own already improves knowledge about safe food handling.

[23] When incentives are introduced in module 3 there is an upward jump in performance in all conditions (see Figures A.1 and A.2; t-tests, $p \leq 0.001$).

[24] A measure of overconfidence can be constructed by matching the self-judgement of a subject of the own performance in module 1 in terms of preparing food under hygienic circumstances (5-point Likert scale) with quintiles in IFSA scores across all subjects in module 1. Taking the difference between the rank for the self-judgement and the rank for the actual performance can be interpreted as a negative (positive) score that indicates over(under)-confidence. For example, someone who rates his performance as "excellent" (top quintile) but has a score in the third quintile gets a miscalibration score of -2.

[25] Pooled data (1033 individuals, N=4132): $\beta_{\text{overconf.}} = -0.080$ ($p = 0.009$), $\beta_{\text{overconf.}\times\text{treat}} = 0.019$ ($p = 0.606$); $\beta_{\text{knowledge}} = -0.055$ ($p = 0.764$), $\beta_{\text{knowledge}\times\text{treat}} = -0.194$ ($p = 0.383$); $\beta_{\text{gamer}} = 0.006$ ($p = 0.977$), $\beta_{\text{gamer}\times\text{treat}} = -0.080$ ($p = 0.674$).

6.3.1 Distraction effects

Some recipes are more complex than others because the cat appears and/or the bread drops to the floor.[26] Overall, subjects perform worse in recipes where the cat appears or the bread drops, or both (see Column 1 in Tables A.37 and A.38), and this is driven by the non-reminded actions (see Columns 3 and 4).[27]

To explore how such disturbances might affect the treatment effects, we consider interactions of the cat appearing or bread dropping with the treatment dummies for *Reminder* or *ManyReminders*, respectively. We observe a negative coefficient on the interaction of the cat appearing with the treatment effect of *Reminder* or *ManyReminders* (see Column 1). A plausible explanation is that touching the cat when removing it serves as a 'reminder' of hand-washing and thus makes some reminders for hand-washing in *Reminder* and *ManyReminders* redundant. If this was the case, then hand-washing should be the driver and in *ManyReminders* we should see the same interaction effect when we consider just the hand-washing score as when we consider the score for the entire set of reminded actions. This is indeed what we find (Wald chi-square test; *Feedback study*: $p = 9681$; *No feedback study*: $p = 0.9122$). In contrast, in recipes where the bread drops, we see no clear-cut evidence of the treatment effect being affected. If at all, reminders seem to partially offset the negative effect of the bread dropping on the overall IFSA score.

6.3.2 Single actions: What influences choices?

This section summarizes the analysis of treatment effects at the level of individual actions that is reported in Appendix A.3. First, choices appear to be rather unrelated to the *relative* incentives in our context (note however that subjects *do* react to the presence of incentives; see Footnote 23). Second, subjects respond to visual cues that draw attention to certain types of actions. Overall, these observations are inconsistent with the interpretation that spillovers solely arise because of cost considerations; rather limited attention seems to matter.

Incentivized actions. Each IFSA action carries the same monetary reward. But because actions take different amounts of time to perform, costs differ and therefore the relative incentives vary. If spillovers were purely driven by cost considerations, we would expect crowding-out of the more time-intensive actions in *Reminder* (such as cleaning the surface, washing the utensils, or throwing-out the bread) rather than the less time-intensive ones (such as rinsing the vegetables or checking the temperature). However, we do not observe a clear pattern that would fit this explanation (see Tables A.8 and A.9).

[26] Appearance of the cat and the bread dropping are randomized across recipes, with the total number of occurrences being fixed for a sequence of four recipes (a level) in modules 3 and 4.

[27] Here we exclude the score for throwing out the dropped bread, which otherwise would mechanically increase performance compared to recipes without the bread dropping.

Actions associated with a strong cue. We observe that subjects almost always remove the cat (see Table A.8) and that there are no treatment differences in whether or not subjects remove the cat in module 3.[28] Note that this action is not incentivized and food preparation can be continued without removing the cat. Effort and time cost considerations thus speak against removing the cat. Yet, the cat leaves a trail of cat hair – a strong visual cue that can prompt people to take action.

Similarly, almost all subjects throw-out the bread after it dropped (see Table A.8) and there are no treatment differences in behavior in module 3.[29] Throwing-out the bread, taking and cutting a new one is rather time consuming. Consequently, pure effort/time cost considerations would speak for pursuing another action instead (given that subjects typically do not perform all actions correctly). Yet, the dropped bread again comes with a strong visual cue — cat hair on the bread – that serves as a prompt to throw it out.

Overall, these findings suggest that subjects pursue actions that are easy to remember rather than the actions that promise the highest reward per time unit.

7 Conclusion

We employ a novel experimental approach based on a computer game to study the effectiveness of reminder nudges on behavior in a complex environment, where a decision maker has to pay attention to and perform multiple, different kinds of actions within a short period of time.

Our results illustrate the importance of considering the contemporaneous and intertemporal spillover effects of nudges. We find that a reminder has a direct positive effect on the reminded action. Yet, reminders have negative spillover effects on non-targeted actions – and this in an environment where we might have expected also some positive spillovers to occur. Overall performance still improves in response to a reminder because the direct impact on the reminded action is larger than the spillovers. Further, we observe that the effects can be scaled in the sense that increasing the number of reminded categories leads to an improvement in overall performance, even though the spillover effects are also more pronounced. We also document intertemporal spillovers to behavior when the nudge is no longer present. While the positive effect on targeted behaviors fades out after withdrawal of the reminders, the negative effect on non-targeted behaviors is persistent. In our setting, treatment with reminders therefore has no net effect on overall performance after reminders are withdrawn.

The game-based approach that our paper introduces allows researchers to precisely administer nudges and collect fine-grained data on behavior in a way that would be infeasible or too

[28] *Reminder vs. Control*: *feedback study* $p = 0.233$, *no feedback study* $p = 0.686$; *ManyReminders vs. Control*: *feedback study* $p = 0.269$, *no feedback study* $p = 0.183$.

[29] *Reminder vs. Control*: *feedback study* $p = 0.197$, *no feedback study* $p = 0.677$, ; *ManyReminders vs. Control*: *feedback study* $p = 0.788$, *no feedback study* $p = 0.970$.

costly in most complex environments. It also has direct practical relevance to the emerging practice of using virtual reality (VR)-based training for complex production processes and safety training. While people can acquire knowledge or skills through on-site or on-the-job training and off-site or classroom-based training, they often face challenges in transferring this into real-life situations. Limited effectiveness and costs of these types of training have led to criticism (e.g., Goulding et al., 2012) and has lead many organizations to experiment with VR-based training (e.g., Dhalmahapatra et al., 2022). Such an environment has many parallels with our setting in that one can implement reminder nudges during the training to repeat particular, especially crucial actions. Our results suggest that this may be a more effective way of establishing routines in targeted behavior than training administered without reminders. But it also highlights the importance of testing the performance in both the targeted and non-targeted behaviors to ensure that training leads to appropriate behavior when people face the tasks for real and reminders no longer are present.

References

Allcott, H. and J. B. Kessler (2018): "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons," *Forthcoming, American Economic Journal: Applied Economics*.

Allcott, H. and T. Rogers (2014): "The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation," *American Economic Review*, 104, 3003–37.

Altmann, S., A. Grunewald, and J. Radbruch (2022): "Interventions and Cognitive Spillovers," *Review of Economic Studies*, 89, 2293–2328.

Archsmith, J. E., A. Heyes, M. J. Neidell, and B. N. Sampat (2021): "The Dynamics of Inattention in the (Baseball) Field," Tech. rep., National Bureau of Economic Research.

Betsch, T., S. Lindow, C. Engel, C. Ulshöfer, and J. Kleber (2015): "Has the world changed? My neighbor might know: effects of social context on routine deviation," *Journal of Behavioral Decision Making*, 28, 50–66.

Brandon, A., J. A. List, R. D. Metcalfe, M. K. Price, and F. Rundhammer (2019): "Testing for crowd out in social nudges: Evidence from a natural field experiment in the market for electricity," *Proceedings of the National Academy of Sciences*, 116, 5293–5298.

Burlig, F., L. Preonas, and M. Woerman (2020): "Panel data and experimental design," *Journal of Development Economics*, 144, 102458.

Byrne, D. P., L. Goette, L. A. Martin, A. Miles, A. Jones, S. Schob, T. Staake, and V. Tiefenbeck (2022): "The habit forming effects of feedback: Evidence from a large-scale field experiment," *SSRN Discussion Paper No. 3974371*.

Cairns, J. and R. Slonim (2011): "Substitution effects across charitable donations," *Economics Letters*, 111, 173–175.

Calzolari, G. and M. Nardotto (2017): "Effective reminders," *Management Science*, 63, 2915–2932.

Caplin, A. and D. Martin (2017): "Defaults and attention: The drop out effect," *Revue économique*, 68, 747–755.

Carlin, B. I., S. Gervais, and G. Manso (2013): "Libertarian paternalism, information production, and financial decision making," *Review of Financial Studies*, 26, 2204–2228.

Castro, J. F., D. Velásquez, A. Beltrán, and G. Yamada (2020): "Spillovers and Long-Run Effects of Messages on Tax Compliance: Experimental Evidence from Peru," Tech. rep., SSRN Discussion Paper No. 3755391.

Cattaneo, M. D., P. Cheung, X. Ma, and Y. Masatlioglu (2021): "Attention Overload," *arXiv preprint arXiv:2110.10650*.

Damgaard, M. T. and C. Gravert (2018): "The hidden costs of nudging: Experimental evidence from reminders in fundraising," *Journal of Public Economics*, 157, 15–26.

de Haan, T. and L. Jona (2018): "'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement," *Economic Journal*, 128, 1180–1206.

De Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and bounding experimenter demand," *American Economic Review*, 108, 3266–3302.

DellaVigna, S. and E. Linos (2022): "RCTs to scale: Comprehensive evidence from two nudge units," *Econometrica*, 90, 81–116.

Dewey-Mattia, D., K. Manikonda, A. J. Hall, M. E. Wise, and S. J. Crowe (2018): "Surveillance for foodborne disease outbreaks—United States, 2009–2015," *MMWR Surveillance Summaries*, 67, 1.

Dhalmahapatra, K., S. Das, and J. Maiti (2022): "On accident causation models, safety training and virtual reality," *International Journal of Occupational Safety and Ergonomics*, 28, 28–44.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual risk attitudes: Measurement, determinants, and behavioral consequences," *Journal of the European Economic Association*, 9, 522–550.

Dolan, P. and M. M. Galizzi (2015): "Like ripples on a pond: behavioral spillovers and their implications for research and policy," *Journal of Economic Psychology*, 47, 1–16.

Duarte, F. and J. S. Hastings (2012): "Fettered consumers and sophisticated firms: evidence from Mexico's privatized social security market," National Bureau of Economic Research Working Paper.

EFSA and ECDC (2018): "The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017," *EFSA Journal*, 16, e05500.

Ek, C. (2017): "Some causes are more equal than others? The effect of similarity on substitution in charitable giving," *Journal of Economic Behavior & Organization*, 136, 45–62.

——— (2018): "Prosocial behavior and policy spillovers: A multi-activity approach," *Journal of Economic Behavior & Organization*, 149, 356–371.

——— (2020): "ANCOVA power calculation in the presence of serial correlation and time shocks: A comment on Burlig et al.(2020)," *Working Paper, University of Gothenburg.*

Ek, C. and J. Miliute-Plepiene (2018): "Behavioral spillovers from food-waste collection in Swedish municipalities," *Journal of Environmental Economics and Management*, 89, 168–186.

Embi, P. J. and A. C. Leonard (2012): "Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study," *Journal of the American Medical Informatics Association*, 19, e145–e148.

Ericson, K. M. (2017): "On the interaction of memory and procrastination: Implications for reminders, deadlines, and empirical estimation," *Journal of the European Economic Association*, 15, 692–719.

Farhi, E. and X. Gabaix (2020): "Optimal taxation with behavioral agents," *American Economic Review*, 110, 298–336.

Filiz-Ozbay, E. and N. Uler (2019): "Demand for giving to multiple charities: An experimental study," *Journal of the European Economic Association*, 17, 725–753.

Frey, E. and T. Rogers (2014): "Persistence: How treatment effects persist after interventions stop," *Policy Insights from the Behavioral and Brain Sciences*, 1, 172–179.

Gabaix, X. (2019): "Behavioral inattention," in *Handbook of Behavioral Economics: Applications and Foundations 1*, Elsevier, vol. 2, 261–343.

Gardner, B. (2015): "A review and analysis of the use of 'habit'in understanding, predicting and influencing health-related behaviour," *Health Psychology Review*, 9, 277–295.

Gilbert, S. J., A. Boldt, C. Sachdeva, C. Scarampi, and P.-C. Tsai (2023): "Outsourcing memory to external tools: A review of 'intention offloading'," *Psychonomic Bulletin & Review*, 30, 60–76.

Goulding, J., W. Nadim, P. Petridis, and M. Alshawi (2012): "Construction industry offsite production: A virtual reality interactive training environment prototype," *Advanced Engineering Informatics*, 26, 103–116.

Grubb, M. D. (2014): "Consumer inattention and bill-shock regulation," *The Review of Economic Studies*, 82, 219–257.

Grubb, M. D. and M. Osborne (2014): "Biased beliefs, learning, and bill shock," *American Economic Review*, 105, 234–271.

Hall, J. D. and J. M. Madsen (2022): "Can behavioral interventions be too salient? Evidence from traffic safety messages," *Science*, 376, eabm3427.

Handel, B. R. (2013): "Adverse selection and inertia in health insurance markets: When nudging hurts," *American Economic Review*, 103, 2643–2682.

Heidhues, P., J. Johnen, and B. Kőszegi (2021): "Browsing versus studying: A pro-market case for regulation," *Review of Economic Studies*, 88, 708–729.

Hummel, D. and A. Maedche (2019): "How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies," *Journal of Behavioral and Experimental Economics*, 80, 47–58.

Hussam, R. and D. Oh (2022): "Negative behavioral transmission," Working Paper, Harvard Business School.

Hussam, R., A. Rabbani, G. Reggiani, and N. Rigol (2022): "Rational habit formation: experimental evidence from handwashing in India," *American Economic Journal: Applied Economics*, 14, 1–41.

Jessoe, K., G. E. Lade, F. Loge, and E. Spang (2021): "Spillovers from behavioral interventions: Experimental evidence from water and energy use," *Journal of the Association of Environmental and Resource Economists*, 8, 315–346.

Jimenez-Gomez, D. (2018): "Nudging and phishing: A theory of behavioral welfare economics," Tech. rep.

Kahana, M. J. (2020): "Computational models of memory search," *Annual Review of Psychology*, 71, 107–138.

Kamenica, E., S. Mullainathan, and R. Thaler (2011): "Helping consumers know themselves," *American Economic Review*, 101, 417–422.

Koch, A. K., D. Mønster, J. Nafziger, and N. Veflen (2022): "Fostering safe food handling among consumers: Causal evidence on game-and video-based online interventions," *Food Control*, 108825.

Lacetera, N., M. Macis, and R. Slonim (2012): "Will there be blood? Incentives and displacement effects in pro-social behavior," *American Economic Journal: Economic Policy*, 4, 186–223.

Malmendier, U. and J. A. Wachter (2021): "Memory of past experiences and economic decisions," SSRN Working Paper No. 4013583.

McKenzie, D. (2012): "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 99, 210–221.

Medina, P. C. (2021): "Side effects of nudging: Evidence from a randomized intervention in the credit card market," *The Review of Financial Studies*, 34, 2580–2607.

Meer, J. (2017): "Does fundraising create new giving?" *Journal of Public Economics*, 145, 82–93.

Merritt, A. C., D. A. Effron, and B. Monin (2010): "Moral self-licensing: When being good frees us to be bad," *Social and Personality Psychology Compass*, 4, 344–357.

Nafziger, J. (2020): "Spillover effects of nudges," *Economic Letters*, 190, 109086.

Neisser, U. and R. Becklen (1975): "Selective looking: Attending to visually specified events," *Cognitive psychology*, 7, 480–494.

Null, C. (2011): "Warm glow, information, and inefficient charitable giving," *Journal of Public Economics*, 95, 455–465.

Redmond, E. C. and C. J. Griffith (2003): "Consumer food handling in the home: a review of food safety studies," *Journal of Food Protection*, 66, 130–161.

Reinstein, D. A. (2011): "Does one charitable contribution come at the expense of another?" *The BE Journal of Economic Analysis & Policy*, 11.

Risko, E. F. and S. J. Gilbert (2016): "Cognitive offloading," *Trends in Cognitive Sciences*, 20, 676–688.

Scallan, E., R. M. Hoekstra, F. J. Angulo, R. V. Tauxe, M.-A. Widdowson, S. L. Roy, J. L. Jones, and P. M. Griffin (2011): "Foodborne illness acquired in the United States—major pathogens," *Emerging Infectious Diseases*, 17, 7.

Seres, G., A. Balleyer, N. Cerutti, J. Friedrichsen, and M. Süer (2021): "Face mask use and physical distancing before and after mandatory masking: No evidence on risk compensation in public waiting lines," *Journal of Economic Behavior & Organization*, 192, 765–781.

Sherif, R. (2021): "Are pro-environment behaviours substitutes or complements? Evidence from the field," Working paper, Max Planck Institute for Tax Law and Public Finance.

Spiegler, R. (2015): "On the equilibrium effects of nudging," *The Journal of Legal Studies*, 44, 389–416.

Sunstein, C. R. (2017): "Nudges that fail," *Behavioural Public Policy*, 1, 4–25.

Taubinsky, D. (2013): "From intentions to actions: A model and experimental evidence of inattentive choice," Working Paper, UC Berkeley.

Taylor, R. L. (2020): "A Mixed Bag: The Hidden Time Costs of Regulating Consumer Behavior," *Journal of the Association of Environmental and Resource Economists*, 7, 345–378.

Thaler, R. (1985): "Mental accounting and consumer choice," *Marketing science*, 4, 199–214.

Thaler, R. H. and C. R. Sunstein (2008): *Nudge: Improving decisions about health, wealth, and happiness*, Penguin.

Thunström, L. (2019): "Welfare effects of nudges: The emotional tax of calorie menu labeling," *Judgment and Decision Making*, 14, 11.

Thunström, L., B. Gilbert, and C. J. Ritten (2018): "Nudges that hurt those already hurting–distributional and unintended effects of salience nudges," *Journal of Economic Behavior & Organization*, 153, 267–282.

Tiefenbeck, V., T. Staake, K. Roth, and O. Sachs (2013): "For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign," *Energy Policy*, 57, 160–171.

Trachtman, H. (2022): "Does promoting one behavior distract from others? Evidence from a field experiment," mimeo, Hebrew University.

Wood, W. and D. T. Neal (2007): "A new look at habits and the habit-goal interface." *Psychological Review*, 114, 843.

——— (2009): "The habitual consumer," *Journal of Consumer Psychology*, 19, 579–592.

Online appendix (for publication)

A.1 Control variables

To to capture individual heterogeneity (cf. eq. 1), the basic set of controls includes the pre-treatment average of the IFSA score per recipe for the seven recipes in modules 1 and 2 ($\bar{Y}_{i,Pre}$). Because subjects complete recipes without being given specific instructions in module 1, play reflects preferences and prior knowledge related to food preparation[30], as well as learning how to play the game. The difference in the average score per recipe in modules 2 and 1 provides us with the measure *Scorediff12* that captures learning and the impact of the information about safe food handling provided in the video prior to the start of module 2 on game play.

As a control for prior knowledge we use the aggregate measure of food safety related knowledge from the survey after module 1: we ask questions about food safety related knowledge and behaviors targeted by the video and the game. 10 questions measure the belief of a subject that a particular action will affect the likelihood of getting foodborne illness (see Table A.3). We designed the game and video interventions to make people aware that certain actions, such as, for example, rinsing chicken, increase the likelihood of getting foodborne illness. Twenty-two questions measure the extent to which participants already perform food safety related behaviors that are targeted by the video and the game, such as whether a subject rinses certain fruits and vegetables (see Table A.4). For the knowledge and behaviors, whenever relevant, items are recoded so that a higher value indicates 'better' knowledge or behavior. We then standardize individual items and aggregate items for the respective groups of outcome measures by taking the average over the individual standardized measures to obtain the control variables for food safety related *PriorKnowledge* and *PriorBehaviors*.

The survey further provides us with a measure of preferences for hygiene behavior and beliefs about how well the player did in the game in terms of hygiene behavior. Specifically, in the survey, we first ask subjects whether they perceived the game as fun, in which dimensions of the game they wanted to perform well (e.g., finishing quickly or preparing the dish hygienically) and how well they think they performed in these dimensions. This provides us the control variables: *Enjoyment*, capturing whether the game is perceived as fun (5-point Likert scale); and *SpeedPref*, capturing how important it was to be quick to complete a recipe (5-point Likert scale).

Further, we ask about individual characteristics, socioeconomic background, risk preferences in general and specific to the food safety domain, food, cooking and hygiene[31] preferences and prior experiences, from which we obtain the other control variables listed in Table A.1.

[30] It is likely that subjects are knowledgeable about some food safety related actions at this stage, but not all (see the evidence in Redmond and Griffith, 2003).

[31] Based on the responses to another question, we can check for consistency with the preferences stated for the game context.

Table A.1: Control variables

Variable	Description	Values
Scorediff12	Difference in the average score per recipe in modules 2 and 1	
Enjoyment	The game is fun	$1, \ldots, 5^a$
SpeedPref	Important to be quick to complete a recipe in the game	$1, \ldots, 5^a$
HygienePref	Importance of preparing food under hygienic circumstances	$1, \ldots, 5^b$
PriorKnowledge	Food safety related knowledge	continuous
PriorBehaviors	Food safety related behaviors	continuous
Age	Age in years	$18, \ldots, 60$
Gender	base category: male, dummy for female, dummy for non-binary/third gender	
SingleHousehold	Dummy=1 if the participant lives in a single-person household	0,1
CurrentSituation	University/college is baseline, dummies for high-school, vocational training, employed in public sector, employed in private sector, self-employed, unemployed, other category	
FreqMeat	How often the participant prepares a warm lunch or dinner with meat (including poultry) on average	
FreqComputerGames	Frequency of playing computer games	$0, \ldots, 7^c$
WorkedFoodSector	Dummy for whether the participant has ever worked in the food industry or in gastronomy/food service, coded 1 if yes and 0 if no	0,1
HealthSector	Dummy for whether the participant has ever worked as a health professional (health worker, nurse, doctor, physician, nutritionist, ...); include health students and health vocational training, coded 1 if yes and 0 if no	0,1
HadFoodPoison	Dummy for whether the participant has ever had food poisoning, coded 1 if yes and 0 if no/don't know	0,1
SharePreprepared	Share of pre-prepared meals that just need to be heated	$0, \ldots, 5^d$
Risk tolerance	Question of Dohmen et al. (2011)	$1, \ldots, 11$

Notes: [a] 1=strongly disagree, 2=somewhat disagree, 3=neither agree nor disagree, 4=somewhat agree, 5=strongly agree [b] 1=very unimportant, 2=slightly unimportant, 3=neutral, 4=slightly important, 5=very important [c] 1=never, 2=0 - 1 hour a month, 3=1 - 3 hours a month, 4=1 - 2 hours a week, 5=4 - 6 hours a week, 6=1 - 2 hours a day, 7=More than 2 hours a day
[d] 1=none (0%), 2=1 in 4 (25%), 3=half (50%), 4=3 in 4 (75%), 5=all (100%)

Table A.2: Extended set of control variables

Variable	Description	Values
InfoSeek	Information seeking about food safety in week prior to experiment	$1, \ldots, 5^a$
Children	Dummy=1 if living with children	0,1
Stressed	How often the participant felt stressed when cooking because of time pressure	$1, \ldots, 5^a$
ConcernedFoodPois	Food-related risk tolerance	$0, \ldots, 11^b$
KitchenCleanPref	Importance of not messing up the kitchen when cooking	$1, \ldots, 5^c$
AvoidWastePref	Importance of avoiding waste	$1, \ldots, 5^c$
HygieneMyth	Agreement with statement "Too much hygiene in the kitchen is the cause of allergies and prevents building up a good immune system"	$1, \ldots, 5^d$

Notes: [a] 1=never, 2=once, 3=twice, 4=3-4 times, 5=5 times or more [b] Concerned about getting sick from food poisoning? 1=not at all concerned ... 11=very concerned [c] 1=very unimportant, 2=slightly unimportant, 3=neutral, 4=slightly important, 5=very important [d] 1=strongly disagree, 2=somewhat disagree, 3=neither agree nor disagree, 4=somewhat agree, 5=strongly agree

Table A.3: Items in *PriorKnowledge*

Description	Reverse coded
Peeling unwashed vegetables/fruit	Yes
Rinsing unwashed vegetables/fruit	No
Picking up within 5 seconds any food that has fallen to the ground	Yes
Heating hamburger meat such that only the inside is pink	Yes
Cooking chicken to an inside temperature of 63 degrees Celsius	Yes
Rinsing a whole chicken before preparation	Yes
Rinsing hands under running water without using soap	Yes
Washing hands with soap under running water	No
Washing cutting boards and kitchen tools in between preparing different food items	No
Rinsing a whole melon	No

Scale: Increases risk by a (1) large (2) small amount, Has no effect on risk (3), Decreases risk by a (3) small (4) large amount

Table A.4: Items in *PriorBehaviors*

Description	Recoded
Targeted behavior 1-4 (Scale 1)	
Did you wash your hands with soap?	No
Did you clean the kitchen surface?	No
Did you rinse a piece of raw meat?	No
Did you clean the cutting board/use a new one?	No
Targeted behavior 5-6 (Scale 2)	
I used a food thermometer	No
I did not check whether the meat is done	Yes[a]
Targeted behavior 7-22 (Scale 3)	
A whole raw chicken	Yes[b]
Raw chicken breasts	Yes[b]
Raw beef	Yes[b]
A whole lettuce	No
A whole watermelon	No
An apple	No
A mango	No
An eggplant	No
An onion	No
String beans	No
Brussels sprouts	No
Potatoes	No
Carrots	No
Berries	No
An avocado	No
Bean sprouts	No

Scale 1: Never (1), Once (2), Twice (3), 3-4 times (4), 5 times or more (5). Scale 2: Yes (1), No (2). Scale 3: How likely would you be to rinse before further preparation/consumption? No chance or almost no chance (1 in 100) (1) ... Certain or practically certain (99 in 100) (11). [a] Recoded 0=Yes, 1=No. [b] Reverse coded.

A.2 Feedback as a confound

In the feedback study, from module 3 onward subjects receive feedback on their performance in the game after each level – a set of four recipes. To check whether feedback is a confounding factor for the results related to Hypotheses 3-4, we examine whether receiving feedback has a differential impact on subsequent performance in the different treatments. Table A.21 reports regressions of the difference between the IFSA scores across levels of module 3 on a treatment dummy.

For *Reminder* vs. *Control*, the gap in the overall IFSA score shrinks when progressing from the first to the second level and from the second to the third level, but the effect is only significant for the change between the first and second levels of module 3. The effect is driven by hand-washing – the reminded action in *Reminder* – for which the gap between *Reminder* and *Control* shrinks from level to level.

For *ManyReminders* vs. *Control* a similar pattern emerges. The gap in the overall IFSA score shrinks from level to level and this is again being driven by the reminded actions.

Doing the same analysis for the no feedback study serves as a placebo test. Table A.22 indeed shows no significant treatment effects for the changes in the overall IFSA score between the first and second levels or between the second and third levels in module 3. Note that this occurs even though ceiling effects could produce negative treatment effects for reminded actions even in the absence of feedback. Specifically, if reminders prompt action, subjects in the treatment condition already do the reminded actions, whereas there still is scope for improved performance in the comparison treatment where these actions are not reminded. Indeed, we see some evidence for this in columns (4), (7), and (8) of Table A.22.

Overall, these results suggest that subjects in *Control* get closer to the performance in the reminder treatments because they learn from feedback and the feedback helps them pay attention to the correct actions. That is, feedback may to some extent be a substitute for reminders and hence make treatment differences become smaller over time in the feedback study. Consequently, the reminders may appear ineffective after their removal even if they would be effective in the absence of feedback. This was the motivation for additionally running the no feedback study.

To further examine the impact of feedback, we look at behavior in the third level of module 3, where subjects have twice received feedback. Performance in *Reminder* is still better than in *Control*, both when considering all actions (with borderline significance) and when considering the reminded action only (see Panel A in Table A.23). Yet, the coefficients for the treatment effects become smaller compared to those in the first level in Table 4, where subjects have not yet received feedback. According to a Wald chi-square test for equality of coefficients across models, the coefficients are significantly different for all actions ($p = 0.0062$) and for hand-washing – the category reminded in *Reminder* ($p < 0.0001$). In contrast, the size of the

negative spillovers is the same as in in the first module: the coefficients for the non-reminded actions are not significantly different ($p = 0.6476$).

Again the no feedback study serves as a placebo test (see Panel B in Table A.23). Here the treatment effect on performance still is present at the end of module 3 and, as expected, coefficients in the third level are not significantly different compared to those in the first level for any of the categories ($p = 0.8848/0.8149/0.9475$ for all actions/reminded/non-reminded).

The comparison of *ManyReminders* and *Control* yields similar insights (see Panel A in Table A.24). All treatment effects observed in the first level of module 3 remain significant in the third level of module 3. The coefficients on the treatment dummy *ManyReminders* are smaller for all actions ($p < 0.0001$) and for the actions reminded in *ManyReminders* ($p < 0.0001$) compared to those in level 1 (see Table A.7). The coefficients for the non-reminded actions are not significantly different ($p = 0.7261$), which means that the size of the negative spillovers remains constant.

A.3 Individual actions

Our main outcome variable aggregates the correctly applied important food safety actions. We now examine on which specific actions the reminder has a positive or negative effect, i.e., whether there is crowding-in or crowding-out of specific actions. In Tables A.8-A.9 we report the immediate effects and in Tables A.10-A.12 the effects after withdrawal of the reminder for *Reminder* vs. *Control*, *ManyReminders* vs. *Control*, and *ManyReminders* vs. *Reminder*, respectively. For space reasons, we only report the specification with the extended set of control variables. For the immediate effects, we pool the feedback and no feedback study to gain power (noting that the latter is a replication of the former). For the effects after the withdrawal of the reminders, we report results for the two studies separately. The results should be interpreted with some caution because the effects are small and we are performing multiple exploratory tests.

A.3.1 Direct effects

The hand-washing reminder is effective in both *Reminder* and *ManyReminders* at the different points (hand-washing in the beginning, after the preparation of meat/vegetables) (see Table A.8). Moreover, we observe no significant difference between *Reminder* and *ManyReminders* in hand-washing behavior at the beginning and after the handling of vegetables, but subjects wash their hands less often in *ManyReminders* than in *Reminder* after meat (see Table A.9).

In *ManyReminders*, the reminder on cleaning the surface is much more effective after the preparation of vegetables than after the preparation of meat and we observe that subjects clean much more often the surface after preparing vegetables in *ManyReminders* than in *Reminder*.

We discuss an explanation for these results in Section 6.1.2. After handling raw meat, three reminders are given at the same time in *ManyReminders* (check the temperature of the meat, wash hands, clean the surface), while subjects in *Reminder* only receive a single reminder at this point (wash hands). The "clean the surface"-reminder is shown last in the reminder list after the meat preparation. Thus, subjects might start with washing hands and checking the temperature, but forget about cleaning the surface then. In contrast, after/when preparing the vegetables, the reminders on washing the hands and cleaning the surface are separated in time, so that only a single reminder appears. Such a single reminder appears to be more effective than a reminder for multiple actions.

Tables A.10 - A.12 show the corresponding effects after the withdrawal of the reminders. We observe in both *Reminder* and *ManyReminders* that subjects in the feedback study wash their hands less often before starting to cook than in *Control*, though this does not replicate the no feedback study. In contrast, they retain the hand-washing habit after the preparation of vegetables and in the feedback study also after preparing meat. One possible interpretation is that the latter two hand-washing actions are associated with a clear cue (touching the ingredient), while hand-washing in the beginning is not. A missing cue could make the formation of habits more difficult. Similarly, subjects continue to clean the surface after preparing vegetables and meat in *ManyReminders* relative to *Control*, but there is no longer a difference when it comes to checking the temperature of the meat.

A.3.2 Indirect effects

Crowding-out might be more likely to occur for actions that take longer (such as cleaning the surface, washing the cutting board and knife, throwing-out and replacing the bread) compared to actions that take less time (such as rinsing vegetables or checking the meat temperature). For *Reminder* vs. *Control* and for *ManyReminders* vs. *Control* (see Table A.8), crowding-out is prominent on the more time-consuming action of cleaning the tools, but appears also on the less time-consuming action of rinsing the vegetables. Overall, this indicates that relative incentives do not appear to be driving the treatment effects.

The comparison between *ManyReminders* and *Reminders* in Table A.9 shows that subjects clean tools less often after preparing meat in *ManyReminders* than in *Reminders*. But there is no significant difference between the two treatments after preparing vegetables. In real-life, cleaning tools after handling meat is more important for food safety than after handling vegetables. The fact that many reminders crowd-out actions that are more important (in real life), suggests that spill-over effects may even be more costly in a real-life setting, where IFSAs do not count equally but have differential impact on food safety.

After the removal of the reminders, we observe in Tables A.10-A.12 that the negative spillover effects discussed above are mainly persistent, though not always statistically significant due to the sample being split into feedback study and no feedback study.

A.4 Alternative outcome measure

In the calculation of the IFSA score, for the action "cleaning the surface" , we check whether a subject cleaned one area around the cutting board. In the information video, subjects were instructed to clean not only the area below the cutting board, but also the areas to the left and right of it. Yet, due to a programming mistake, subjects in the feedback study got a feedback star for only cleaning one area. Thus, we built the IFSA score on whether a subject cleaned one area, and check for robustness here.

Checking whether the subject cleans one area indicates whether the respective reminder induces attention and leads to some action. Checking whether the subject cleans all three areas indicates whether the respective reminder does not only induce attention, but also whether it gets subjects to remember how to correctly perform the action – or look up how to do so in the video or help – and then perform the correct action. We report the results for the main hypotheses in Tables A.26-A.29. The results for the comparisons of *Reminder* vs. *Control* are robust. Unsurprisingly, as one of the reminders is about cleaning the surface, the *ManyReminders* intervention is less effective, but the comparisons between *ManyReminders* and *Control* are robust; yet there no longer is a significant difference between *ManyReminders* compared to *Reminder*. This indicates that the surface reminder induces subjects to take action, but does not induce additional effort to check the information video.

A.5 Accessing information about food safety

We find no evidence of a treatment effect on how often subjects accessed food safety information (outside of the study information) during the period of the study, how often subjects clicked to reopen any videos, and in how much time they spent on the videos (two-sided t-tests). In the feedback study, there is no difference in information seeking (1.24 times in *Control*, 1.24 times in *Reminder*, and 1.21 times in *ManyReminders*; *Reminder* vs. *Control*: $p = 0.9090$, N=508; *ManyReminders* vs. *Control*: $p = 0.5384$, N=506; *ManyReminders* vs. *Reminder*:$p = 0.6558$, N=508), reopening videos (0.48 times in *Control*, 0.64 times in *Reminder*, and 0.65 times in *ManyReminders*; *Reminder* vs. *Control*: $p = 0.3715$; *ManyReminders* vs. *Control*: $p = 0.1352$; *ManyReminders* vs. *Reminder*: $p = 0.9710$; the duration of videos being open (149 seconds in *Control*, 151 seconds in *Reminder*, and 156 seconds in *ManyReminders*; *Reminder* vs. *Control*: $p = 0.6658$; *ManyReminders* vs. *Control*: $p = 0.4748$; *ManyReminders* vs. *Reminder*: $p = 0.6436$). The same patterns hold for the no feedback study.

A.6 Handling vegetables vs. handling meat

We use as dependent variables the difference between the first four recipes of module 4 and the first four recipes of module 3 in hand-washing (surface cleaning) after handling meat and after handling vegetables, respectively. Because hand-washing is reminded in both treatments, we run regressions for both *Reminder* vs. em Control and *ManyReminders* vs. *Control*. For surface cleaning, we consider only *ManyReminders* vs. *Control*. The coefficients on the respective treatment dummies are significantly different for performance after meat and after vegetables in four out of the six regressions.[32]

A.7 Food preparation order

Some food preparation orders are safer than others. In the information video on food safety, we tell about appropriate hygiene behavior for preventing cross-contamination, but do not suggest a specific order in which actions should be taken. An appropriate order is first implicitly suggested in the video play-through at the start of module 3 (bread-vegetable-chicken). This sequence limits potential cross-contamination from the most risky category – raw meat to the other food items and cross-contamination from possibly unwashed raw vegetables to bread. The recipe on the fridge suggests the reverse order (see Figure 1). The 'safe' preparation order takes slightly longer than the one where one prepares the chicken first (noting that the chicken needs time to be cooked, during which one can do other things). The fact that 58 percent of the subjects apply the safe preparation order in module 1, where they are not yet aware that the study has to do with food safety, suggests that subjects apply their real world habits. After seeing the food safety information video at the start of module 2, 73 percent of the subjects apply the safe preparation order. In module 3, 78 percent in *Control*, 76 percent in *Reminder*, and 76 percent in *ManyReminders* do so. The latter observation suggests that subjects also take non-incentivized food safety actions despite higher costs. In this case, it is not the visual cue that makes them take the action, but real world habits.

[32] Hand-washing after vegetables compared to after meat (Wald chi-square test for equality of coefficients across models): (i) *Reminder* vs. *Control*: feedback study $p < 0.0001$, no feedback study $p = 0.9137$; (ii) *ManyReminders* vs. *Control*: feedback study $p < 0.0001$, no feedback study $p = 0.4744$. Surface cleaning after vegetables compared to after meat, *ManyReminders* vs. *Control*: feedback study $p < 0.0001$, no feedback study $p < 0.0001$.

A.8 Additional tables

Table A.5: Balance table (feedback study)

Variable	*Control*	*Reminder*	*ManyReminders*	*Reminder* vs. *Control*	*ManyReminder* vs. *Control*	*Reminder* vs. *ManyReminders*
Scorediff12	2.740	2.898	2.882	0.158	0.142	0.016
	(2.288)	(2.320)	(2.137)	(0.204)	(0.196)	(0.198)
Enjoyment	3.812	3.796	3.776	-0.016	-0.036	0.020
	(0.958)	(0.979)	(0.929)	(0.086)	(0.084)	(0.085)
SpeedPref	4.306	4.227	4.209	-0.078	-0.097	0.019
	(0.842)	(0.876)	(0.880)	(0.076)	(0.076)	(0.078)
HygienePref	4.392	4.486	4.610	0.094	0.218***	-0.124**
	(0.834)	(0.736)	(0.630)	(0.070)	(0.066)	(0.061)
PriorKnowledge	-0.036	0.032	0.033	0.068**	0.069**	-0.001
	(0.379)	(0.397)	(0.382)	(0.034)	(0.034)	(0.035)
PriorBehaviors	-0.025	-0.002	0.061	0.022	0.085**	-0.063
	(0.432)	(0.429)	(0.447)	(0.038)	(0.039)	(0.039)
Age	33.600	33.851	33.240	0.251	-0.360	0.611
	(10.627)	(10.784)	(10.727)	(0.948)	(0.946)	(0.953)
Female	0.482	0.475	0.488	-0.008	0.006	-0.014
	(0.501)	(0.500)	(0.501)	(0.044)	(0.044)	(0.044)
Nonbinary	0.008	0.012	0.008	0.004	0.000	0.004
	(0.088)	(0.108)	(0.089)	(0.009)	(0.008)	(0.009)
SingleHousehold	0.173	0.149	0.114	-0.024	-0.058*	0.035
	(0.379)	(0.357)	(0.319)	(0.033)	(0.031)	(0.030)
Highschool	0.008	0.004	0.008	-0.004	0.000	-0.004
	(0.088)	(0.063)	(0.089)	(0.007)	(0.008)	(0.007)
Vocational	0.004	0.004	0.012	-0.000	0.008	-0.008
	(0.063)	(0.063)	(0.108)	(0.006)	(0.008)	(0.008)
Publicsector	0.188	0.208	0.228	0.020	0.040	-0.021
	(0.392)	(0.407)	(0.421)	(0.035)	(0.036)	(0.037)
Privatesector	0.314	0.329	0.339	0.016	0.025	-0.009
	(0.465)	(0.471)	(0.474)	(0.041)	(0.042)	(0.042)
Selfemployed	0.118	0.114	0.110	-0.004	-0.007	0.003
	(0.323)	(0.318)	(0.314)	(0.028)	(0.028)	(0.028)
Unemployed	0.161	0.122	0.102	-0.039	-0.058*	0.019
	(0.368)	(0.327)	(0.304)	(0.031)	(0.030)	(0.028)
Other	0.012	0.055	0.016	0.043***	0.004	0.039**
	(0.108)	(0.228)	(0.125)	(0.016)	(0.010)	(0.016)

FreqMeat	5.035	5.051	5.130	0.016	0.095	-0.079
	(1.525)	(1.527)	(1.518)	(0.135)	(0.135)	(0.135)
FreqComputerGames	3.890	4.059	4.063	0.169	0.173	-0.004
	(1.963)	(2.060)	(2.063)	(0.178)	(0.179)	(0.183)
WorkedFoodSector	0.224	0.286	0.260	0.063	0.036	0.026
	(0.417)	(0.453)	(0.439)	(0.039)	(0.038)	(0.040)
HealthSector	0.035	0.008	0.024	-0.027**	-0.012	-0.016
	(0.185)	(0.088)	(0.152)	(0.013)	(0.015)	(0.011)
HadFoodPoison	0.486	0.525	0.512	0.039	0.026	0.014
	(0.501)	(0.500)	(0.501)	(0.044)	(0.044)	(0.044)
SharePreprepared	2.106	2.078	2.067	-0.027	-0.039	0.012
	(1.031)	(0.940)	(0.982)	(0.087)	(0.089)	(0.085)
Risktolerance	4.992	4.729	4.795	-0.263	-0.197	-0.066
	(2.204)	(2.213)	(2.079)	(0.196)	(0.190)	(0.190)
Observations	255	255	254	510	509	509

Notes: Mean (std.dev.). T-tests: * p<0.10, ** p<0.05, *** p<0.01. Variables are explained in Tables A.1 and A.2.

Table A.6: Balance table (no feedback study)

Variable	*Control*	*Reminder*	*ManyReminders*	*Reminder* vs. *Control*	*ManyReminder* vs. *Control*	*Reminder* vs. *ManyReminders*
Scorediff12	2.901	2.988	2.652	0.087	-0.249	0.336*
	(2.295)	(2.313)	(2.237)	(0.202)	(0.201)	(0.199)
Enjoyment	3.833	3.937	3.808	0.104	-0.025	0.129
	(0.934)	(0.943)	(0.951)	(0.082)	(0.084)	(0.083)
SpeedPref	4.262	4.262	4.357	0.000	0.095	-0.095
	(0.858)	(0.804)	(0.848)	(0.073)	(0.076)	(0.072)
HygienePref	4.611	4.590	4.541	-0.021	-0.070	0.049
	(0.668)	(0.698)	(0.632)	(0.060)	(0.058)	(0.058)
PriorKnowledge	-0.008	0.011	-0.028	0.019	-0.020	0.038
	(0.388)	(0.375)	(0.397)	(0.033)	(0.035)	(0.034)
PriorBehaviors	-0.003	0.000	-0.023	0.003	-0.020	0.023
	(0.428)	(0.445)	(0.407)	(0.038)	(0.037)	(0.037)
Age	36.075	35.469	35.620	-0.607	-0.456	-0.151
	(9.985)	(9.714)	(10.190)	(0.862)	(0.896)	(0.868)
Female	0.460	0.480	0.467	0.019	0.006	0.013
	(0.499)	(0.501)	(0.500)	(0.044)	(0.044)	(0.044)
Nonbinary	0.000	0.015	0.008	0.015*	0.008	0.007
	(0.000)	(0.121)	(0.088)	(0.008)	(0.006)	(0.009)
SingleHousehold	0.159	0.140	0.114	-0.019	-0.045	0.026

	(0.366)	(0.348)	(0.318)	(0.031)	(0.030)	(0.029)
Highschool	0.004	0.000	0.004	-0.004	-0.000	-0.004
	(0.063)	(0.000)	(0.063)	(0.004)	(0.006)	(0.004)
Vocational	0.004	0.000	0.008	-0.004	0.004	-0.008
	(0.063)	(0.000)	(0.088)	(0.004)	(0.007)	(0.005)
Publicsector	0.250	0.258	0.255	0.008	0.005	0.003
	(0.434)	(0.439)	(0.437)	(0.038)	(0.039)	(0.038)
Privatesector	0.397	0.354	0.369	-0.043	-0.028	-0.014
	(0.490)	(0.479)	(0.483)	(0.042)	(0.043)	(0.042)
Selfemployed	0.095	0.114	0.090	0.019	-0.005	0.024
	(0.294)	(0.319)	(0.287)	(0.027)	(0.026)	(0.027)
Unemployed	0.135	0.129	0.165	-0.006	0.030	-0.036
	(0.342)	(0.336)	(0.372)	(0.030)	(0.032)	(0.031)
Other	0.016	0.026	0.031	0.010	0.015	-0.006
	(0.125)	(0.159)	(0.175)	(0.013)	(0.014)	(0.015)
FreqMeat	5.048	5.085	5.090	0.037	0.043	-0.005
	(1.501)	(1.558)	(1.520)	(0.134)	(0.134)	(0.134)
FreqComputerGames	3.825	4.133	3.980	0.307*	0.155	0.152
	(1.933)	(1.893)	(2.017)	(0.167)	(0.175)	(0.170)
WorkedFoodSector	0.222	0.295	0.251	0.073*	0.029	0.044
	(0.417)	(0.457)	(0.434)	(0.038)	(0.038)	(0.039)
HealthSector	0.004	0.015	0.008	0.011	0.004	0.007
	(0.063)	(0.121)	(0.088)	(0.009)	(0.007)	(0.009)
HadFoodPoison	0.563	0.513	0.471	-0.051	-0.093**	0.042
	(0.497)	(0.501)	(0.500)	(0.044)	(0.044)	(0.044)
SharePreprepared	2.032	1.985	1.980	-0.047	-0.051	0.005
	(0.944)	(0.939)	(0.933)	(0.082)	(0.083)	(0.082)
Risktolerance	4.937	4.779	4.973	-0.158	0.036	-0.194
	(2.134)	(2.418)	(2.189)	(0.200)	(0.192)	(0.202)
Observations	252	271	255	523	507	526

Notes: Mean (std.dev.). T-tests: * p<0.10, ** p<0.05, *** p<0.01. Variables are explained in Tables A.1 and A.2.

Table A.7: Secondary Hypothesis 1: *ManyReminders* vs. *Control*

Panel A: Feedback study (509 individuals; N=2036)									
	All actions			Reminded actions in *ManyReminders*			Non-reminded actions in *ManyReminders*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	1.33***	1.35***	1.38***	1.61***	1.64***	1.67***	-0.29***	-0.29***	-0.28***
vs. Control	(0.15)	(0.14)	(0.14)	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)	(0.10)
Adj. R2	0.53	0.59	0.61	0.48	0.52	0.53	0.40	0.47	0.48
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.16	8.16	8.16	3.18	3.18	3.18	4.98	4.98	4.98
Std.dev.	2.91	2.91	2.91	1.63	1.63	1.63	1.67	1.67	1.67

Panel B: No feedback study (507 individuals; N=2028)									
	All actions			Reminded actions in *ManyReminders*			Non-reminded actions in *ManyReminders*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	1.02***	1.04***	1.06***	1.53***	1.54***	1.55***	-0.51***	-0.50***	-0.50***
vs. Control	(0.15)	(0.14)	(0.14)	(0.09)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)
Adj. R2	0.52	0.58	0.58	0.45	0.49	0.49	0.45	0.49	0.49
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.48	8.48	8.48	3.35	3.35	3.35	5.13	5.13	5.13
Std.dev.	2.78	2.78	2.78	1.59	1.59	1.59	1.56	1.56	1.56

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.8: Single Actions (pooled data)

Panel A: Reminder vs. Control (1033 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
	0.26***	0.44***	0.27***	-0.00	-0.10**	-0.07*	-0.04*	-0.02	-0.02	0.00	0.00	-0.03	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)
N	4132	4132	4132	4132	4132	4132	4132	4132	4132	2066	2065	4132	4132
Adj. R2	0.22	0.32	0.22	0.09	0.29	0.33	0.18	0.26	0.13	0.04	0.01	0.09	0.18
Mean (Control)	0.60	0.42	0.64	0.82	0.78	1.50	0.22	0.61	0.77	0.97	0.99	0.77	0.84
	0.49	0.49	0.48	0.38	0.97	0.85	0.42	0.49	0.42	0.17	0.11	0.42	0.36

Panel B: ManyReminders vs. Control (1016 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
	0.28***	0.43***	0.16***	-0.03*	-0.15***	-0.21***	0.54***	0.12***	0.06***	0.00	0.00	-0.04	-0.03**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)	(0.04)	(0.02)	(0.02)	(0.02)	(0.01)	(0.00)	(0.02)	(0.02)
N	4064	4064	4064	4063	4063	4064	4064	4064	4064	2032	2030	4064	4064
Adj. R2	0.22	0.30	0.18	0.09	0.28	0.33	0.39	0.21	0.10	0.06	0.02	0.09	0.19
Mean (Control)	0.60	0.42	0.64	0.82	0.78	1.50	0.22	0.61	0.77	0.97	0.99	0.77	0.84
	0.49	0.49	0.48	0.38	0.97	0.85	0.42	0.49	0.42	0.17	0.11	0.42	0.36

Notes: Regressions are based on the pooled data for the feedback study and the no feedback study. Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: Score in for the category listed in the column heading. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control or ManyReminders vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* or *ManyReminders*, respectively, and 0 if in *Control*. Specifications include a recipe fixed effect, a control for the average pre-treatment score in modules 1 and 2 and the extended set of control variables. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (treatment) is the mean of the dependent variable in *Control*. The categories of actions reminded in *Reminder* are hand-washing and in *ManyReminders* they are hand-washing, surface cleaning, and checking the temperature of the meat. Only in about half of the recipes does bread drop and/or the cat appear.

Table A.9: Single Actions (pooled data)

ManyReminders vs. Reminder (1035 individuals)													
	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
	0.00	-0.01	-0.11***	-0.03	-0.04	-0.13***	0.58***	0.14***	0.08***	0.00	-0.00	0.00	-0.02
	(0.01)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)
N	4140	4140	4140	4139	4139	4140	4140	4140	4140	2070	2069	4140	4140
Adj. R2	0.07	0.08	0.10	0.09	0.30	0.38	0.41	0.23	0.13	0.05	0.02	0.10	0.19
Controls	Ext	Ext	Ext	Ext	Ext	Ext	Ext	Ext	Ext	Ext	Ext	Ext	Ext
Mean (Reminder)	0.88	0.87	0.93	0.83	0.71	1.48	0.19	0.61	0.77	0.97	0.99	0.75	0.83
	0.32	0.33	0.26	0.38	0.95	0.86	0.40	0.49	0.42	0.17	0.09	0.44	0.38

Notes: Regressions are based on the pooled data for the feedback study and the no feedback study. Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: Score for the category listed in the column heading. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include a recipe fixed effect, a control for the average pre-treatment score in modules 1 and 2 and the extended set of control variables. Standard errors are clustered at the individual level (in parentheses). * $p<0.10$, ** $p<0.05$, *** $p<0.01$. Mean (treatment) is the mean of the dependent variable in *Reminder*. The categories of actions reminded in *Reminder* are hand-washing and in *ManyReminders* they are hand-washing, surface cleaning, and checking the temperature of the meat. Only in about half of the recipes does bread drop and/or the cat appear.

Table A.10: Single Actions (after withdrawal of reminders)

Panel A: Reminder vs. Control (feedback study; 510 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
Reminder vs. Control	-0.12***	0.16***	0.13***	-0.03	-0.10	-0.08	0.00	0.00	-0.04*	-0.01	0.00	-0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.02)	(0.07)	(0.05)	(0.03)	(0.03)	(0.02)	(0.01)	(0.01)	(0.03)	(0.02)
N	2040	2040	2040	2040	2040	2040	2040	2040	2040	1020	1018	2040	2040
Adj. R2	0.11	0.16	0.12	0.05	0.29	0.20	0.22	0.20	0.07	0.04	0.01	0.08	0.18
Mean (Control)	0.79	0.49	0.66	0.87	0.89	1.62	0.29	0.62	0.91	0.99	0.99	0.77	0.90
Std.dev.	0.41	0.50	0.47	0.34	0.99	0.77	0.45	0.49	0.28	0.12	0.11	0.42	0.30

Panel B: Reminder vs. Control (no feedback study; 523 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
Reminder vs. Control	-0.02	0.24***	0.08***	-0.04	-0.03	-0.15***	-0.04	-0.06*	-0.05*	-0.00	-0.00	0.00	0.03*
	(0.03)	(0.03)	(0.03)	(0.03)	(0.07)	(0.06)	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.03)	(0.02)
N	2092	2092	2092	2092	2092	2092	2092	2092	2092	1046	1046	2092	2092
Adj. R2	0.15	0.23	0.14	0.11	0.26	0.24	0.18	0.26	0.12	0.08	0.03	0.07	0.18
Mean (Control)	0.63	0.41	0.68	0.81	0.66	1.62	0.21	0.64	0.80	0.97	0.99	0.82	0.87
Std.dev.	0.48	0.49	0.47	0.40	0.93	0.78	0.41	0.48	0.40	0.18	0.08	0.38	0.33

Notes: Data are from the first four recipes of module 4. Dependent variable: Score in for the category listed in the column heading. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control or Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* or *Reminder*, respectively, and 0 if in *Control*. Specifications include a recipe fixed effect, a control for the average pre-treatment score in modules 1 and 2 and the extended set of control variables. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (treatment) is the mean of the dependent variable in *Control*. The categories of actions reminded in *Reminder* are hand-washing. Only in about half of the recipes does bread drop and/or the cat appear.

Table A.11: Single Actions (after withdrawal of reminders)

Panel A: ManyReminders vs. Control (feedback study; 509 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
ManyReminders vs. Control	-0.15***	0.16***	0.06*	-0.03	-0.12*	-0.18***	0.18***	0.08**	-0.01	0.00	-0.00	-0.04	0.02
	(0.03)	(0.03)	(0.03)	(0.02)	(0.07)	(0.06)	(0.03)	(0.03)	(0.02)	(0.01)	(0.01)	(0.03)	(0.02)
N	2036	2036	2036	2036	2036	2036	2036	2036	2036	1018	1017	2036	2036
Adj. R2	0.13	0.17	0.10	0.05	0.27	0.28	0.19	0.19	0.05	0.05	0.03	0.11	0.20
Mean (Control)	0.79	0.49	0.66	0.87	0.89	1.62	0.29	0.62	0.91	0.99	0.99	0.77	0.90
Std.dev.	0.41	0.50	0.47	0.34	0.99	0.77	0.45	0.49	0.28	0.12	0.11	0.42	0.30

Panel B: ManyReminders vs. Control (no feedback study; 507 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
ManyReminders vs. Control	0.01	0.22***	-0.00	-0.01	-0.11*	-0.26***	0.21***	0.06*	0.03	-0.00	-0.01	-0.05	0.03
	(0.03)	(0.03)	(0.03)	(0.03)	(0.06)	(0.06)	(0.03)	(0.03)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)
N	2028	2028	2028	2028	2028	2028	2028	2028	2028	1014	1014	2028	2028
Adj. R2	0.18	0.24	0.19	0.12	0.24	0.28	0.20	0.23	0.11	0.09	0.03	0.07	0.18
Mean (Control)	0.63	0.41	0.68	0.81	0.66	1.62	0.21	0.64	0.80	0.97	0.99	0.82	0.87
Std.dev.	0.48	0.49	0.47	0.40	0.93	0.78	0.41	0.48	0.40	0.18	0.08	0.38	0.33

Notes: Data are from the first four recipes of module 4. Dependent variable: Score in for the category listed in the column heading. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control or ManyReminders vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* or *ManyReminders*, respectively, and 0 if in *Control*. Specifications include a recipe fixed effect, a control for the average pre-treatment score in modules 1 and 2 and the extended set of control variables. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (treatment) is the mean of the dependent variable in *Control*. The categories of actions reminded *ManyReminders* are hand-washing, surface cleaning, and checking the temperature of the meat. Only in about half of the recipes does bread drop and/or the cat appear.

Table A.12: Single Actions (after withdrawal of reminders)

Panel A: ManyReminders vs. Reminder (feedback study; 509 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
ManyReminders vs. Reminder	-0.05	0.00	-0.07**	-0.00	-0.02	-0.09*	0.17***	0.06*	0.02	0.01	-0.01	-0.02	0.02
	(0.03)	(0.03)	(0.03)	(0.02)	(0.07)	(0.06)	(0.03)	(0.03)	(0.02)	(0.01)	(0.01)	(0.04)	(0.02)
N	2036	2036	2036	2036	2036	2036	2036	2036	2036	1018	1017	2036	2036
Adj. R2	0.11	0.11	0.06	0.06	0.31	0.28	0.20	0.18	0.08	0.06	0.04	0.06	0.16
Mean (Reminder)	0.68	0.66	0.81	0.84	0.82	1.59	0.31	0.66	0.89	0.98		0.77	0.90
Std.dev.	0.47	0.47	0.39	0.37	0.98	0.80	0.46	0.48	0.31	0.15	0.10	0.42	0.30

Panel B: ManyReminders vs. Reminder (no feedback study; 526 individuals)

	Wash Hands			Rinse	Clean Tools After		Clean Surface After		Check	Throw out	Remove	Safe	Keep
	In Beginning	After Veggies	After Meat	Veggies	Veggies	Meat	Veggies	Meat	Temperature	Bread	Cat	Order	Time
ManyReminders vs. Reminder	0.01	-0.02	-0.09***	0.02	-0.08	-0.10*	0.24***	0.12***	0.09***	-0.00	-0.01	-0.04	-0.00
	(0.03)	(0.03)	(0.03)	(0.03)	(0.06)	(0.06)	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.03)	(0.02)
N	2104	2104	2104	2104	2104	2104	2104	2104	2104	1052	1052	2104	2104
Adj. R2	0.12	0.15	0.15	0.12	0.28	0.31	0.21	0.28	0.15	0.08	0.03	0.08	0.16
Mean (Reminder)	0.62	0.66	0.79	0.79	0.63	1.51	0.18	0.60	0.77	0.97		0.82	0.90
Std.dev.	0.48	0.47	0.41	0.41	0.92	0.85	0.39	0.49	0.42	0.17	0.07	0.39	0.30

Notes: Data are from the first four recipes of module 4. Dependent variable: Score in for the category listed in the column heading. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control or ManyReminders vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* or *ManyReminders*, respectively, and 0 if in *Control*. Specifications include a recipe fixed effect, a control for the average pre-treatment score in modules 1 and 2 and the extended set of control variables. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (treatment) is the mean of the dependent variable in *Control*. The categories of actions reminded in *Reminder* are hand-washing and in *ManyReminders* they are hand-washing, surface cleaning, and checking the temperature of the meat. Only in about half of the recipes does bread drop and/or the cat appear.

Table A.13: Secondary Hypothesis 2: Direct and spillover effects

Panel A: Feedback study (509 individuals; N=2036)									
	Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.09	-0.08	-0.07	0.75***	0.76***	0.78***	-0.17*	-0.19**	-0.20**
vs. Reminder	(0.06)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)	(0.10)	(0.09)	(0.09)
Adj. R2	0.10	0.12	0.12	0.33	0.35	0.37	0.44	0.50	0.50
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.63	2.63	2.63	4.15	4.15	4.15	4.99	4.99	4.99
Std.dev.	0.78	0.78	0.78	1.29	1.29	1.29	1.64	1.64	1.64

Panel B: No feedback study (526 individuals; N=2104)									
	Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.16***	-0.15***	-0.15***	0.57***	0.62***	0.62***	-0.21**	-0.19**	-0.20**
vs. Reminder	(0.05)	(0.05)	(0.05)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)
Adj. R2	0.11	0.15	0.15	0.30	0.36	0.36	0.48	0.54	0.54
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.73	2.73	2.73	4.34	4.34	4.34	4.93	4.93	4.93
Std.dev.	0.66	0.66	0.66	1.24	1.24	1.24	1.59	1.59	1.59

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) hand-washing – the category of actions reminded in *Reminder*; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.14: Secondary Hypothesis 4: Direct and spillover effects

Panel A: Feedback study (509 individuals; N=2036)									
	Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.12*	-0.12*	-0.12*	0.10	0.11	0.13	-0.14	-0.13	-0.11
vs. Reminder	(0.06)	(0.06)	(0.06)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)
Adj. R2	0.15	0.16	0.16	0.26	0.28	0.29	0.38	0.42	0.43
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.16	2.16	2.16	4.02	4.02	4.02	5.23	5.23	5.23
Std.dev.	0.92	0.92	0.92	1.49	1.49	1.49	1.55	1.55	1.55

Panel B: No feedback study (526 individuals; N=2104)									
	Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.10	-0.09	-0.10	0.32***	0.36***	0.36***	-0.18*	-0.18*	-0.18*
vs. Reminder	(0.07)	(0.07)	(0.07)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Adj. R2	0.20	0.24	0.24	0.33	0.38	0.39	0.42	0.46	0.47
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.07	2.07	2.07	3.63	3.63	3.63	4.88	4.88	4.88
Std.dev.	0.96	0.96	0.96	1.54	1.54	1.54	1.60	1.60	1.60

Notes: Data are from first four recipes of module 4 in the feedback study. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: (1)-(3) hand-washing – the category of actions reminded in *Reminder*; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.15: Hypothesis 1 (all recipes in module 3)

Panel A: Feedback study (510 individuals; N=6119)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.50***	0.52***	0.50***	0.71***	0.70***	0.70***	-0.21*	-0.18	-0.19*
vs. Control	(0.15)	(0.14)	(0.14)	(0.05)	(0.05)	(0.05)	(0.12)	(0.12)	(0.12)
Adj. R2	0.46	0.52	0.53	0.29	0.32	0.32	0.44	0.51	0.52
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.83	8.83	8.83	1.90	1.90	1.90	6.93	6.93	6.93
Std.dev.	2.88	2.88	2.88	1.00	1.00	1.00	2.24	2.24	2.24
Panel B: No feedback study (523 individuals; N=6275)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.63***	0.60***	0.64***	1.02***	1.01***	1.01***	-0.39***	-0.41***	-0.37***
vs. Control	(0.15)	(0.13)	(0.13)	(0.06)	(0.05)	(0.05)	(0.12)	(0.11)	(0.11)
Adj. R2	0.47	0.55	0.56	0.40	0.43	0.44	0.45	0.52	0.53
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.43	8.43	8.43	1.65	1.65	1.65	6.77	6.77	6.77
Std.dev.	2.82	2.82	2.82	1.03	1.03	1.03	2.15	2.15	2.15

Notes: Data are from all 12 recipes of module 3. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except hand-washing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.16: Secondary Hypothesis 1: *ManyReminders* vs. *Control* (all recipes in module 3)

Panel A: Feedback study (509 individuals; N=6107)									
	All actions			Reminded actions in *ManyReminders*			Non-reminded actions in *ManyReminders*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	0.87***	0.91***	0.95***	1.22***	1.25***	1.27***	-0.35***	-0.34***	-0.32***
vs. Control	(0.15)	(0.14)	(0.14)	(0.09)	(0.08)	(0.08)	(0.10)	(0.09)	(0.09)
Adj. R2	0.43	0.49	0.50	0.38	0.42	0.43	0.35	0.42	0.43
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.83	8.83	8.83	3.64	3.64	3.64	5.19	5.19	5.19
Std.dev.	2.88	2.88	2.88	1.60	1.60	1.60	1.62	1.62	1.62

Panel B: No feedback study (507 individuals; N=6084)									
	All actions			Reminded actions in *ManyReminders*			Non-reminded actions in *ManyReminders*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	0.91***	0.95***	0.96***	1.40***	1.43***	1.44***	-0.49***	-0.48***	-0.48***
vs. Control	(0.16)	(0.14)	(0.14)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Adj. R2	0.47	0.53	0.54	0.40	0.46	0.47	0.42	0.47	0.47
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.43	8.43	8.43	3.35	3.35	3.35	5.08	5.08	5.08
Std.dev.	2.82	2.82	2.82	1.61	1.61	1.61	1.55	1.55	1.55

Notes: Data are from all 12 recipes of module 3. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.17: Hypothesis 2 (all recipes in module 3)

Panel A: Feedback study (509 individuals; N=6108)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	0.36**	0.37***	0.41***
	(0.15)	(0.14)	(0.13)
Adj. R2	0.41	0.47	0.48
Controls	No	Basic	Ext
Mean (Control)	9.55	9.55	9.55
Std.dev.	2.74	2.74	2.74

Panel B: No feedback study (526 individuals; N=6311)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	0.25*	0.33**	0.32**
	(0.14)	(0.13)	(0.13)
Adj. R2	0.48	0.55	0.56
Controls	No	Basic	Ext
Mean (Control)	9.21	9.21	9.21
Std.dev.	2.57	2.57	2.57

Notes: Data are from all 12 recipes of module 3. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: Overall IFSA score. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.18: Secondary Hypothesis 2: Direct and spillover effects (all recipes in module 3)

Panel A: Feedback study (509 individuals; N=6108)									
	Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.09**	-0.08*	-0.07	0.54***	0.56***	0.58***	-0.18*	-0.19**	-0.18**
vs. Reminder	(0.05)	(0.05)	(0.05)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)
Adj. R2	0.10	0.13	0.13	0.26	0.29	0.31	0.39	0.45	0.45
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.65	2.65	2.65	4.41	4.41	4.41	5.13	5.13	5.13
Std.dev.	0.77	0.77	0.77	1.38	1.38	1.38	1.66	1.66	1.66

Panel B: No feedback study (526 individuals; N=6311)									
	Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.21***	-0.20***	-0.20***	0.47***	0.53***	0.53***	-0.22**	-0.20**	-0.21**
vs. Reminder	(0.05)	(0.05)	(0.05)	(0.09)	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)
Adj. R2	0.15	0.20	0.20	0.29	0.36	0.37	0.45	0.50	0.51
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.71	2.71	2.71	4.32	4.32	4.32	4.90	4.90	4.90
Std.dev.	0.69	0.69	0.69	1.31	1.31	1.31	1.59	1.59	1.59

Notes: Data are from all 12 recipes of module 3. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) hand-washing – the category of actions reminded in *Reminder*; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.19: Secondary Hypothesis 3 (all recipes in module 4)

Panel A: Feedback study (509 individuals; N=4071)

	All actions			Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
ManyReminders vs. Reminder	0.01	0.11	0.14	0.06	0.08	0.08	0.29***	0.34***	0.35***	-0.29***	-0.23**	-0.21**
	(0.18)	(0.16)	(0.16)	(0.06)	(0.06)	(0.06)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)	(0.10)
Adj. R2	0.34	0.41	0.42	0.18	0.22	0.22	0.25	0.31	0.32	0.29	0.36	0.37
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	9.31	9.31	9.31	2.03	2.03	2.03	3.91	3.91	3.91	5.40	5.40	5.40
Std.dev.	2.75	2.75	2.75	0.96	0.96	0.96	1.53	1.53	1.53	1.54	1.54	1.54

Panel B: No feedback study

Module 4 only has four recipes in the no feedback study

Notes: Data are from all eight recipes of module 4. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: (1)-(3) Hand-washing – the category of actions reminded in *Reminder*; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Control) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.20: Hypothesis 4 and Secondary Hypothesis 4 (all recipes in module 4)

Panel A: Feedback study (509 individuals; N=4071)												
	All actions			Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
ManyReminders vs. Reminder	-0.01	0.03	0.09	-0.13**	-0.12**	-0.12**	0.10	0.12	0.15	-0.11	-0.09	-0.06
	(0.17)	(0.16)	(0.16)	(0.06)	(0.06)	(0.06)	(0.10)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)
Adj. R2	0.38	0.43	0.44	0.16	0.19	0.19	0.26	0.29	0.30	0.35	0.40	0.41
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	2.26	2.26	2.26	2.26	2.26	2.26	4.19	4.19	4.19	5.33	5.33	5.33
Std.dev.	0.88	0.88	0.88	0.88	0.88	0.88	1.47	1.47	1.47	1.55	1.55	1.55
Panel B: No feedback study												
Module 4 only has four recipes in the no feedback study												

Notes: Data are from all eight recipes of module 4. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable: (1)-(3) Hand-washing – the category of actions reminded in *Reminder*; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.21: Feedback as a confound (feedback study)

	All actions		Reminded actions in *Reminder*		Non-reminded actions in *Reminder*		Reminded actions in *ManyReminders*		Never-reminded actions	
	Difference in the average score per recipe between module 3 levels									
	1-2	2-3	1-2	2-3	1-2	2-3	1-2	2-3	1-2	2-3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Reminder vs. Control	-0.30***	-0.17	-0.28***	-0.11**	-0.03	-0.06				
	(0.11)	(0.12)	(0.05)	(0.05)	(0.09)	(0.10)				
ManyReminders vs. Control	-0.56***	-0.30**	-0.26***	-0.15***			-0.45***	-0.30***	-0.11	-0.00
	(0.11)	(0.12)	(0.05)	(0.05)			(0.07)	(0.08)	(0.08)	(0.08)
ManyReminders vs. Reminder	-0.25**	-0.13	0.02	-0.04			-0.21***	-0.19**	-0.04	0.06
	(0.11)	(0.12)	(0.05)	(0.05)			(0.07)	(0.08)	(0.08)	(0.08)

Notes: Dependent variable: Difference between the stated levels (a collection of four recipes) in the average per recipe for (1)-(2) the overall IFSA score; (3)-(4) hand-washing – the category of actions reminded in *Reminder*; (5)-(6) all actions except hand-washing; (7)-(8) hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (9)-(10) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (treatment vs. baseline) that is equal to 1 if the subject participated in treatment and 0 if in the stated baseline. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01.

Table A.22: Feedback as a confound (no feedback study, placebo test)

	All actions		Reminded actions in *Reminder*		Non-reminded actions in *Reminder*		Reminded actions in *ManyReminders*		Never-reminded actions	
	Difference in the average score per recipe between module 3 levels									
	1-2	2-3	1-2	2-3	1-2	2-3	1-2	2-3	1-2	2-3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Reminder vs. Control	-0.01	-0.01	-0.00	-0.01	-0.01	-0.00				
	(0.09)	(0.08)	(0.03)	(0.04)	(0.07)	(0.07)				
ManyReminders vs. Control	-0.12	-0.11	-0.05	-0.11**			-0.12**	-0.16***	0.00	0.05
	(0.10)	(0.09)	(0.04)	(0.04)			(0.06)	(0.06)	(0.06)	(0.05)
ManyReminders vs. Reminder	-0.11	-0.10	-0.04	-0.10***			-0.08	-0.17***	-0.03	0.07
	(0.09)	(0.10)	(0.04)	(0.04)			(0.06)	(0.06)	(0.05)	(0.06)

Notes: Dependent variable: Difference between the stated levels (a collection of four recipes) in the average per recipe for (1)-(2) the overall IFSA score; (3)-(4) hand-washing – the category of actions reminded in *Reminder*; (5)-(6) all actions except hand-washing; (7)-(8) hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (9)-(10) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (treatment vs. baseline) that is equal to 1 if the subject participated in treatment and 0 if in the stated baseline. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01.

Table A.23: Hypothesis 1 (last level of module 3)

Panel A: Feedback study (510 individuals; N=2040)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.30	0.35*	0.33*	0.54***	0.55***	0.55***	-0.24	-0.20	-0.21
vs. Control	(0.19)	(0.18)	(0.18)	(0.06)	(0.06)	(0.06)	(0.15)	(0.15)	(0.15)
Adj. R2	0.36	0.43	0.44	0.22	0.27	0.27	0.36	0.43	0.44
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	9.27	9.27	9.27	2.08	2.08	2.08	7.20	7.20	7.20
Std.dev.	2.88	2.88	2.88	0.96	0.96	0.96	2.21	2.21	2.21

Panel B: No feedback study (523 individuals; N=2091)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.64***	0.60***	0.65***	1.02***	1.01***	1.02***	-0.38***	-0.41***	-0.37***
vs. Control	(0.17)	(0.16)	(0.16)	(0.06)	(0.06)	(0.06)	(0.14)	(0.13)	(0.13)
Adj. R2	0.43	0.51	0.52	0.39	0.43	0.44	0.41	0.48	0.49
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.36	8.36	8.36	1.63	1.63	1.63	6.73	6.73	6.73
Std.dev.	2.88	2.88	2.88	1.01	1.01	1.01	2.19	2.19	2.19

Notes: Data are from the last four recipes of module 3. That is, after subjects in the feedback study have received feedback on the IFSA score for two levels (sets of four recipes). Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except hand-washing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.24: Secondary Hypothesis 1: *ManyReminders* vs. *Control* (last level of module 3)

Panel A: Feedback study (509 individuals; N=2036)									
	All actions			Reminded actions in *ManyReminders*			Non-reminded actions in *ManyReminders*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	0.50**	0.56***	0.60***	0.88***	0.90***	0.92***	-0.38***	-0.34***	-0.32***
vs. Control	(0.20)	(0.19)	(0.19)	(0.11)	(0.11)	(0.11)	(0.12)	(0.11)	(0.11)
Adj. R2	0.33	0.39	0.41	0.28	0.32	0.34	0.30	0.37	0.38
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	9.27	9.27	9.27	3.96	3.96	3.96	5.32	5.32	5.32
Std.dev.	2.88	2.88	2.88	1.57	1.57	1.57	1.59	1.59	1.59

Panel B: No feedback study (507 individuals; N=2028)									
	All actions			Reminded actions in *ManyReminders*			Non-reminded actions in *ManyReminders*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	0.80***	0.86***	0.86***	1.25***	1.30***	1.30***	-0.46***	-0.44***	-0.44***
vs. Control	(0.19)	(0.17)	(0.17)	(0.12)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)
Adj. R2	0.43	0.49	0.51	0.36	0.43	0.44	0.40	0.44	0.44
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.36	8.36	8.36	3.33	3.33	3.33	5.03	5.03	5.03
Std.dev.	2.88	2.88	2.88	1.63	1.63	1.63	1.55	1.55	1.55

Notes: Data are from the last four recipes of module 3. That is, after subjects in the feedback study have received feedback on the IFSA score for two levels (sets of four recipes). Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.25: Hypothesis 2 and Secondary Hypothesis 2 (last level of module 3)

Panel A: Feedback study (509 individuals; N=2036)												
	All actions			Reminded actions in *Reminder*			Reminded actions in *ManyReminders*			Never-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
ManyReminders vs. Reminder	0.19	0.20	0.26	-0.11*	-0.10*	-0.10*	0.34***	0.36***	0.39***	-0.15	-0.16	-0.13
	(0.19)	(0.18)	(0.18)	(0.06)	(0.06)	(0.06)	(0.11)	(0.10)	(0.10)	(0.11)	(0.11)	(0.11)
Adj. R2	0.34	0.40	0.41	0.10	0.15	0.15	0.21	0.26	0.28	0.34	0.41	0.41
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	9.78	9.78	9.78	2.66	2.66	2.66	4.58	4.58	4.58	5.20	5.20	5.20
Std.dev.	2.86	2.86	2.86	0.76	0.76	0.76	1.43	1.43	1.43	1.67	1.67	1.67
Panel B: No feedback study (526 individuals; N=2103)												
ManyReminders vs. Reminder	0.15	0.23	0.21	-0.29***	-0.27***	-0.27***	0.34***	0.40***	0.39***	-0.19*	-0.17*	-0.19*
	(0.17)	(0.16)	(0.16)	(0.06)	(0.06)	(0.06)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Adj. R2	0.43	0.52	0.53	0.17	0.25	0.25	0.27	0.36	0.38	0.41	0.47	0.48
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	9.15	9.15	9.15	2.68	2.68	2.68	4.30	4.30	4.30	4.84	4.84	4.84
Std.dev.	2.63	2.63	2.63	0.73	0.73	0.73	1.36	1.36	1.36	1.59	1.59	1.59

Notes: Data are from the last four recipes of module 3. That is, after subjects in the feedback study have received feedback on the IFSA score for two levels (sets of four recipes). Dependent variable: (1)-(3) Hand-washing – the category of actions reminded in *Reminder*; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.26: Hypothesis 1 (alternative outcome measure)

Panel A: Feedback study (510 individuals; N=2040)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.78***	0.80***	0.79***	0.93***	0.92***	0.91***	-0.19	-0.15	-0.16
vs. Control	(0.14)	(0.14)	(0.13)	(0.06)	(0.06)	(0.06)	(0.12)	(0.12)	(0.12)
Adj. R2	0.53	0.59	0.60	0.35	0.37	0.39	0.53	0.59	0.60
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	7.72	7.72	7.72	1.65	1.65	1.65	6.51	6.51	6.51
Std.dev.	2.74	2.74	2.74	1.02	1.02	1.02	2.30	2.30	2.30
Panel B: No feedback study (523 individuals; N=2092)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.73***	0.69***	0.72***	1.03***	1.01***	1.01***	-0.39***	-0.41***	-0.38***
vs. Control	(0.14)	(0.12)	(0.12)	(0.06)	(0.06)	(0.06)	(0.12)	(0.11)	(0.11)
Adj. R2	0.51	0.59	0.60	0.40	0.43	0.44	0.49	0.56	0.57
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	7.94	7.94	7.94	1.67	1.67	1.67	6.81	6.81	6.81
Std.dev.	2.62	2.62	2.62	1.05	1.05	1.05	2.13	2.13	2.13

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable based on alternative measure were three surface areas instead of one need to be cleaned: (1)-(3) Overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except hand-washing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.27: Hypothesis 2 (alternative outcome measure)

Panel A: Feedback study (509 individuals; N=2036)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	0.01	-0.01	0.01
	(0.14)	(0.13)	(0.13)
Adj. R2	0.50	0.57	0.57
Controls	No	Basic	Ext
Mean (Control)	8.72	8.72	8.72
Std.dev.	2.45	2.45	2.45

Panel B: No feedback study (526 individuals; N=2104)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	-0.28**	-0.23*	-0.23*
	(0.13)	(0.12)	(0.12)
Adj. R2	0.53	0.60	0.60
Controls	No	Basic	Ext
Mean (Control)	8.82	8.82	8.82
Std.dev.	2.37	2.37	2.37

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable based on alternative measure were three surface areas instead of one need to be cleaned: Overall IFSA score. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.28: Hypothesis 3 (alternative outcome measure)

Panel A: Feedback study (510 individuals; N=2040)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	-0.03	-0.03	-0.04	0.17***	0.16***	0.17***	-0.25*	-0.24*	-0.25*
vs. Control	(0.16)	(0.15)	(0.15)	(0.06)	(0.06)	(0.06)	(0.14)	(0.13)	(0.13)
Adj. R2	0.37	0.43	0.44	0.17	0.19	0.20	0.36	0.41	0.43
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	8.64	8.64	8.64	1.94	1.94	1.94	7.19	7.19	7.19
Std.dev.	2.52	2.52	2.52	0.96	0.96	0.96	2.08	2.08	2.08
Panel B: No feedback study (523 individuals; N=2092)									
	All actions			Reminded actions			Non-reminded actions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Reminder	0.09	0.05	0.08	0.31***	0.30***	0.30***	-0.37***	-0.39***	-0.37***
vs. Control	(0.15)	(0.15)	(0.15)	(0.06)	(0.06)	(0.06)	(0.13)	(0.12)	(0.13)
Adj. R2	0.44	0.50	0.51	0.23	0.27	0.28	0.43	0.49	0.49
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Control)	7.90	7.90	7.90	1.72	1.72	1.72	6.69	6.69	6.69
Std.dev.	2.49	2.49	2.49	0.96	0.96	0.96	2.05	2.05	2.05

Notes: Data are from the first four recipes of module 4. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. Dependent variable based on alternative measure were three surface areas instead of one need to be cleaned: (1)-(3) Overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except handwashing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.29: Hypothesis 4 (alternative outcome measure)

Panel A: Feedback study (509 individuals; N=2036)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	-0.24	-0.22	-0.18
	(0.16)	(0.15)	(0.16)
Adj. R2	0.42	0.46	0.47
Controls	No	Basic	Ext
Mean (Reminder)	8.80	8.80	8.80
Std.dev.	2.51	2.51	2.51

Panel B: No feedback study (526 individuals; N=2104)			
	(1)	(2)	(3)
ManyReminders vs. Reminder	-0.25	-0.21	-0.22
	(0.16)	(0.15)	(0.15)
Adj. R2	0.48	0.54	0.54
Controls	No	Basic	Ext
Mean (Reminder)	8.13	8.13	8.13
Std.dev.	2.69	2.69	2.69

Notes: Data are from the first four recipes of module 4. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable based on alternative measure were three surface areas instead of one need to be cleaned: Overall IFSA score. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.30: Impact of feedback and reminders (Reminder vs. Control, pooled data)

	All	Reminded	Non-Reminded
	(1)	(2)	(3)
Reminder vs. Control	-0.02	-0.01	-0.02
	(0.11)	(0.05)	(0.09)
Feedback	1.22***	0.47***	0.75***
	(0.14)	(0.06)	(0.11)
Feedback*Reminder	-0.45**	-0.38***	-0.06
	(0.19)	(0.08)	(0.15)
Constant	-0.11	-0.04	-0.07
	(0.08)	(0.04)	(0.06)
Adj. R2	0.10	0.08	0.08
Mean (Control)	0.51	0.19	0.31
Std.dev.	1.66	0.74	1.25

Notes: Regressions are based on the pooled data for the feedback study and the no feedback study. N=1038. Dependent variable: Difference in the average per recipe between the last four recipes and the first four recipes of module 3 for (1) the overall IFSA score; (2) hand-washing – the category of actions reminded in *Reminder*; (3) all actions except hand-washing. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. OLS regressions of the dependent variable on a treatment dummy that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*, a dummy that is equal 1 if the subject participated in the feedback study and 0 if in the no feedback study, and an interaction between the treatment and feedback dummies. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.31: Impact of feedback and reminders (ManyReminders vs. Control, pooled data)

	All actions	Reminded actions in *ManyReminders*	Non-reminded actions in *ManyReminders*
	(1)	(2)	(3)
ManyReminders vs. Control	-0.23*	-0.28***	0.05
	(0.13)	(0.09)	(0.07)
Feedback	1.22***	0.79***	0.43***
	(0.14)	(0.09)	(0.08)
Feedback*ManyReminders	-0.63***	-0.47***	-0.16
	(0.20)	(0.13)	(0.12)
Constant	-0.11	-0.01	-0.09**
	(0.08)	(0.05)	(0.05)
Adj. R2	0.10	0.13	0.03
Mean (Control)	0.51	0.39	0.12
Std.dev.	1.66	1.06	0.95

Notes: Regressions are based on the pooled data for the feedback study and the no feedback study. N=1020. Dependent variable: Difference in the average per recipe between the last four recipes and the first four recipes of module 3 for (1)-(3) the overall IFSA score; (4)-(6) hand-washing – the category of actions reminded in *Reminder*; (7)-(9) all actions except hand-washing. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. OLS regressions of the dependent variable on a treatment dummy that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Control*, a dummy that is equal 1 if the subject participated in the feedback study and 0 if in the no feedback study, and an interaction between the treatment and feedback dummies. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.32: Impact of feedback and reminders (ManyReminders vs. Reminder, pooled data)

	All actions	Reminded actions in *Reminder*	Non-reminded actions in *Reminder*	Reminded actions in *ManyReminders*	Never-reminded actions
	(1)	(2)	(3)	(4)	(5)
ManyReminders vs. Reminder	-0.21	-0.15***	-0.07	-0.25***	0.04
	(0.13)	(0.05)	(0.10)	(0.08)	(0.07)
Feedback	0.77***	0.08	0.69***	0.48***	0.30***
	(0.13)	(0.05)	(0.11)	(0.08)	(0.08)
Feedback*ManyReminders	-0.18	0.12	-0.30*	-0.16	-0.03
	(0.20)	(0.08)	(0.16)	(0.12)	(0.12)
Constant	-0.13*	-0.05*	-0.08	-0.04	-0.09*
	(0.08)	(0.03)	(0.06)	(0.04)	(0.05)
Adj. R2	0.05	0.02	0.05	0.06	0.02
Mean (Reminder)	0.24	-0.01	0.25	0.19	0.05
Std.dev.	1.57	0.62	1.29	0.91	0.97

Notes: Regressions are based on the pooled data for the feedback study and the no feedback study. N=1038. Dependent variable: Difference in the average per recipe between the last four recipes and the first four recipes of module 3 for (1)-(3) the overall IFSA score; (3)-(4) hand-washing – the category of actions reminded in *Reminder*; (5)-(6) all actions except hand-washing; (7)-(8) hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (9)-(10) all actions except the reminded actions in *ManyReminders*. In the feedback study, from module 3 on feedback on the IFSA score is given after each set of four recipes. OLS regressions of the dependent variable on a treatment dummy that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*, a dummy that is equal 1 if the subject participated in the feedback study and 0 if in the no feedback study, and an interaction between the treatment and feedback dummies. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.33: Decreasing returns to more reminders/reminding more categories: Hand-washing

Panel A: Feedback study (511 individuals; N=2043)									
	At recipe start			After handling vegetables			After handling meat		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	0.02	0.02	0.02	-0.01	0.00	0.01	-0.10***	-0.10***	-0.10***
vs. Reminder	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Adj. R2	0.04	0.07	0.07	0.07	0.08	0.09	0.08	0.09	0.09
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	0.86	0.86	0.86	0.85	0.85	0.85	0.91	0.91	0.91
Std.dev.	0.35	0.35	0.35	0.35	0.35	0.35	0.28	0.28	0.28

Panel B: No feedback study (526 individuals; N=2104)									
	At recipe start			After handling vegetables			After handling meat		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ManyReminders	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.12***	-0.12***	-0.12***
vs. Reminder	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
N	2108	2104	2104	2108	2104	2104	2108	2104	2104
Individuals	527	526	526	527	526	526	527	526	526
Adj. R2	0.06	0.10	0.10	0.06	0.08	0.08	0.10	0.12	0.12
Controls	No	Basic	Ext	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	0.90	0.90	0.90	0.89	0.89	0.89	0.94	0.94	0.94
Std.dev.	0.30	0.30	0.30	0.32	0.32	0.32	0.24	0.24	0.24

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) hand-washing at the start of a recipe; (4)-(6) hand-washing after handling vegetables; (7)-(9) hand-washing after handling meat. In *ManyReminders*, the reminder message for hand-washing contains a single reminded action at the start of the recipe and after handling vegetables, but three reminded actions after handling meat. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.34: Decreasing returns to more reminded categories: Cleaning kitchen surfaces

Panel A: Feedback study (511 individuals; N=2043)						
	After handling vegetables			After handling meat		
	(1)	(2)	(3)	(4)	(5)	(6)
ManyReminders	0.61***	0.61***	0.62***	0.17***	0.17***	0.17***
vs. Reminder	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Adj. R2	0.43	0.46	0.47	0.20	0.22	0.23
Controls	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	0.17	0.17	0.17	0.58	0.58	0.58
Std.dev.	0.38	0.38	0.38	0.49	0.49	0.49

Panel B: No feedback study (527 individuals; N=2108)						
	After handling vegetables			After handling meat		
	(1)	(2)	(3)	(4)	(5)	(6)
ManyReminders	0.53***	0.54***	0.54***	0.11***	0.13***	0.13***
vs. Reminder	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Adj. R2	0.33	0.36	0.37	0.21	0.25	0.25
Controls	No	Basic	Ext	No	Basic	Ext
Mean (Reminder)	0.21	0.21	0.21	0.64	0.64	0.64
Std.dev.	0.41	0.41	0.41	0.48	0.48	0.48

Notes: Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) surface cleaning after handling vegetables; (4)-(6) surface cleaning after handling meat. In *ManyReminders*, the reminder message for handwashing contains a single reminded action at the start of the recipe and after handling vegetables, but three reminded actions after handling meat. OLS regressions of the dependent variable on a treatment dummy (ManyReminders vs. Reminder) that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Reminder*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Reminder) is the mean of the dependent variable in *Reminder*.

Table A.35: Hypothesis 1: Potential experimenter demand effects

Panel A: Feedback study						
	All actions		Reminded actions		Non-reminded actions	
	(1)	(2)	(3)	(4)	(5)	(6)
Reminder	0.44**	0.08	0.70***	0.37***	-0.26*	-0.28
vs. Control	(0.18)	(0.22)	(0.09)	(0.12)	(0.15)	(0.19)
N	1396	1108	1396	1108	1396	1108
Individuals	349	277	349	277	349	277
Adj. R2	0.57	0.59	0.27	0.22	0.56	0.58
Mean (Control)	7.88	7.58	1.53	1.41	6.35	6.17
Std.dev.	2.83	2.72	0.98	0.93	2.27	2.23

Panel B: No feedback study						
	All actions		Reminded actions		Non-reminded actions	
	(1)	(2)	(3)	(4)	(5)	(6)
Reminder	0.37**	-0.07	0.82***	0.63***	-0.45***	-0.70***
vs. Control	(0.18)	(0.25)	(0.08)	(0.12)	(0.15)	(0.20)
N	1352	1036	1352	1036	1352	1036
Individuals	338	259	338	259	338	259
Adj. R2	0.59	0.57	0.35	0.30	0.56	0.56
Mean (Control)	8.15	7.92	1.53	1.41	6.62	6.51
Std.dev.	2.68	2.64	1.01	0.97	2.09	2.10

Notes: The first (second) specification for each dependent variable excludes those subjects from the analysis who comply with all reminders in all (more than half) of the recipes. Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(2) Overall IFSA score; (4)-(3) hand-washing – the category of actions reminded in *Reminder*; (6)-(6) all actions except hand-washing. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include the extended set of controls, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard errors are clustered at the individual level (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

Table A.36: Secondary Hypothesis 1: Potential experimenter demand effects

Panel A: Feedback study

	All actions		Reminded actions in *ManyReminders*		Non-reminded actions in *ManyReminders*	
	(1)	(2)	(3)	(4)	(5)	(6)
ManyReminders	1.33***	1.16***	1.65***	1.50***	-0.32***	-0.34***
vs. Control	(0.14)	(0.15)	(0.09)	(0.10)	(0.10)	(0.11)
N	1820	1608	1820	1608	1820	1608
Individuals	455	402	455	402	455	402
Adj. R2	0.57	0.55	0.48	0.44	0.45	0.45
Mean (Control)	8.09	7.97	3.14	3.06	4.95	4.90
Std.dev.	2.88	2.82	1.60	1.56	1.67	1.66

Panel B: No feedback study

	All actions		Reminded actions in *ManyReminders*		Non-reminded actions in *ManyReminders*	
	(1)	(2)	(3)	(4)	(5)	(6)
ManyReminders	1.03***	0.99***	1.50***	1.39***	-0.48***	-0.41***
vs. Control	(0.15)	(0.16)	(0.10)	(0.11)	(0.10)	(0.11)
N	1848	1640	1848	1640	1848	1640
Individuals	462	410	462	410	462	410
Adj. R2	0.54	0.54	0.44	0.42	0.48	0.48
Mean (Control)	8.39	8.30	3.29	3.24	5.10	5.06
Std.dev.	2.74	2.69	1.56	1.54	1.55	1.54

Notes: The first (second) specification for each dependent variable excludes those subjects from the analysis who comply with all reminders in all (more than half) of the recipes. Data are from the first four recipes of module 3. That is, before any feedback on the IFSA score is given in the feedback study. Dependent variable: (1)-(3) Overall IFSA score; (4)-(6) Hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (7)-(9) all actions except the reminded actions in *ManyReminders*. OLS regressions of the dependent variable on a treatment dummy (Reminder vs. Control) that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*. Specifications include, next to the controls indicated in the table, a recipe fixed effect and a control for the average pre-treatment score in modules 1 and 2. Standard

Table A.37: Impact of Additional Disturbances (Reminder vs. Control)

	Feedback study			No feedback study		
	All actions	Reminded actions	Non-reminded actions	All actions	Reminded actions	Non-reminded actions
	(1)	(2)	(3)	(4)	(5)	(6)
Reminder vs. Control	0.53***	0.73***	-0.21*	0.70***	1.11***	-0.40***
	(0.14)	(0.06)	(0.12)	(0.14)	(0.06)	(0.12)
Cat in Recipe	-0.30***	0.11***	-0.42***	0.03	0.25***	-0.22***
	(0.07)	(0.04)	(0.05)	(0.06)	(0.03)	(0.05)
Cat*Reminder	-0.25***	-0.19***	-0.05	-0.22***	-0.27***	0.06
	(0.09)	(0.04)	(0.07)	(0.08)	(0.04)	(0.07)
Bread drops in Recipe	-0.47***	-0.15***	-0.32***	-0.11*	0.01	-0.12**
	(0.06)	(0.03)	(0.04)	(0.06)	(0.03)	(0.05)
Bread*Reminder	0.24***	0.14***	0.11	-0.05	-0.04	-0.01
	(0.09)	(0.04)	(0.07)	(0.07)	(0.04)	(0.06)
Bread and Cat in Recipe	-0.18***	0.06**	-0.23***	-0.00	0.14***	-0.14***
	(0.05)	(0.03)	(0.04)	(0.04)	(0.02)	(0.04)
(Bread & Cat)*Reminder	-0.05	-0.08***	0.04	-0.07	-0.13***	0.06
	(0.06)	(0.03)	(0.05)	(0.06)	(0.03)	(0.05)
Constant	2.37***	0.86***	1.52**	1.50*	-0.05	1.55**
	(0.81)	(0.29)	(0.70)	(0.87)	(0.37)	(0.70)
N	6119	6119	6119	6275	6275	6275
Individuals	510	510	510	523	523	523
Adj. R2	0.53	0.32	0.52	0.56	0.44	0.53
Mean (Control)	7.84	1.90	5.95	7.45	1.65	5.79
Std.dev.	2.86	1.00	2.22	2.79	1.03	2.12

Notes: Data are from all recipes of module 3. Dependent variable: (1) & (4) The overall IFSA score†; (2) & (5) hand-washing – the category of reminded in *Reminder*; (3) & (6) all except hand-washing†. †We exclude the score for throwing out the dropped bread, which otherwise would mechanically increase performance compared to recipes without the bread dropping. In the feedback study, feedback on the IFSA score is given after each level of module 3. OLS regressions of the dependent variable on a treatment dummy that is equal to 1 if the subject participated in treatment *Reminder* and 0 if in *Control*, a dummy that is equal 1 if the subject participated in the feedback study and 0 if in the no feedback study, dummies for whether in a recipe the cat appeared (Cat), the bread dropped (Bread) or both (Cat and Bread; individual Cat/Bread dummies are set to zero for such a recipe), and interactions between the dummies and the treatment *Reminder*. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

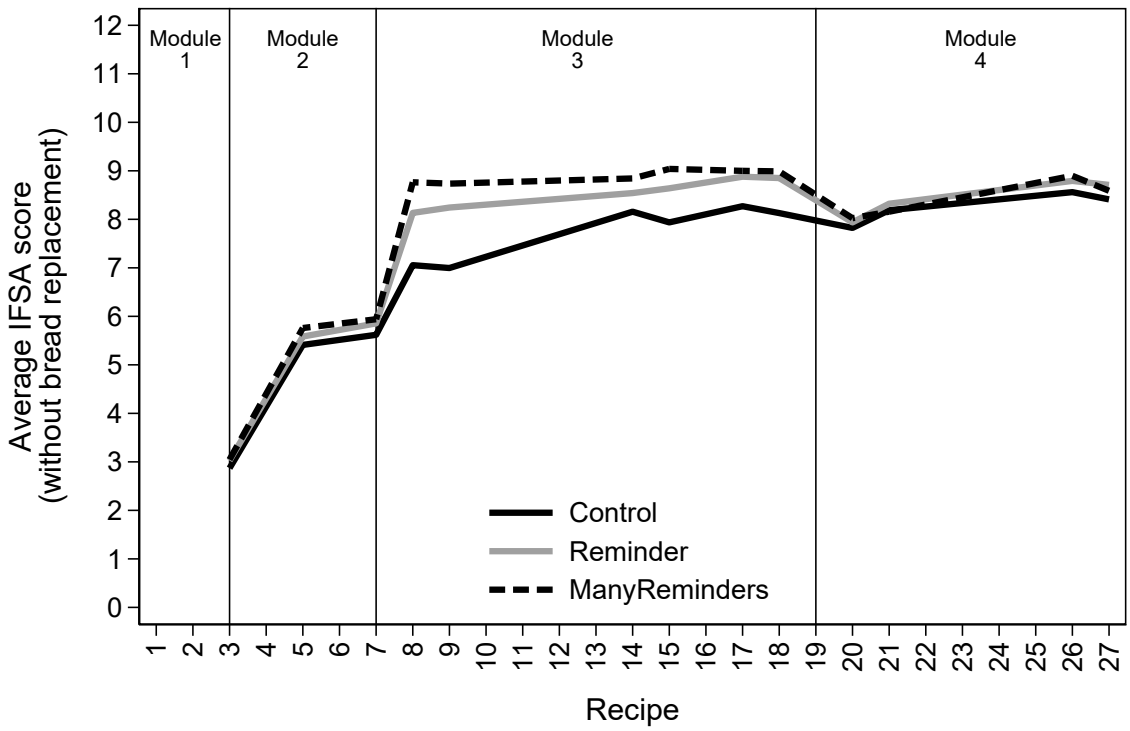
Table A.38: Impact of Additional Disturbances (ManyReminders vs. Control)

	Feedback study			No feedback study		
	All actions	Reminded actions in ManyReminders	Non-reminded actions in ManyReminders	All actions	Reminded actions in ManyReminders	Non-reminded actions in ManyReminders
	(1)	(2)	(3)	(4)	(5)	(6)
ManyReminders vs. Control	0.97***	1.29***	-0.32***	1.01***	1.53***	-0.52***
	(0.14)	(0.08)	(0.09)	(0.15)	(0.10)	(0.10)
Cat in Recipe	-0.25***	-0.04	-0.21***	0.06	0.23***	-0.17***
	(0.07)	(0.05)	(0.04)	(0.06)	(0.04)	(0.04)
Cat*ManyReminders	-0.21**	-0.19***	-0.01	-0.23***	-0.34***	0.11**
	(0.09)	(0.06)	(0.05)	(0.09)	(0.06)	(0.05)
Bread drops in Recipe	-0.45***	-0.28***	-0.18***	-0.09*	-0.04	-0.06
	(0.06)	(0.05)	(0.04)	(0.06)	(0.04)	(0.04)
Bread*ManyReminders	0.19**	0.18***	0.01	0.09	0.04	0.05
	(0.08)	(0.06)	(0.05)	(0.08)	(0.06)	(0.05)
Bread and Cat in Recipe	-0.16***	-0.03	-0.13***	0.01	0.08***	-0.08***
	(0.05)	(0.03)	(0.03)	(0.05)	(0.03)	(0.03)
(Bread & Cat)*ManyReminders	-0.05	-0.05	-0.00	-0.08	-0.14***	0.05
	(0.07)	(0.05)	(0.04)	(0.07)	(0.05)	(0.04)
Constant	1.76**	0.95**	0.81	2.24**	0.82	1.42**
	(0.79)	(0.48)	(0.55)	(1.00)	(0.61)	(0.60)
N	6107	6107	6107	6084	6084	6084
Individuals	509	509	509	507	507	507
Adj. R2	0.51	0.43	0.43	0.54	0.47	0.46
Mean (Control)	7.84	3.64	4.21	7.45	3.35	4.10
Std.dev.	2.86	1.60	1.60	2.79	1.61	1.52

Notes: Data are from all recipes of module 3. Dependent variable: (1) & (4) The overall IFSA score†; (2) & (5) hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders*; (3)-(6) all actions except the reminded actions in *ManyReminders*†. †We exclude the score for throwing out the dropped bread, which otherwise would mechanically increase performance compared to recipes without the bread dropping. In the feedback study, feedback on the IFSA score is given after each level of module 3. OLS regressions of the dependent variable on a treatment dummy that is equal to 1 if the subject participated in treatment *ManyReminders* and 0 if in *Control*, a dummy that is equal 1 if the subject participated in the feedback study and 0 if in the no feedback study, dummies for whether in a recipe the cat appeared (Cat), the bread dropped (Bread) or both (Cat and Bread; individual Cat/Bread dummies are set to zero for such a recipe), and interactions between the dummies and the treatment *ManyReminders*. Robust standard errors (in parentheses). * p<0.10, ** p<0.05, *** p<0.01. Mean (Control) is the mean of the dependent variable in *Control*.

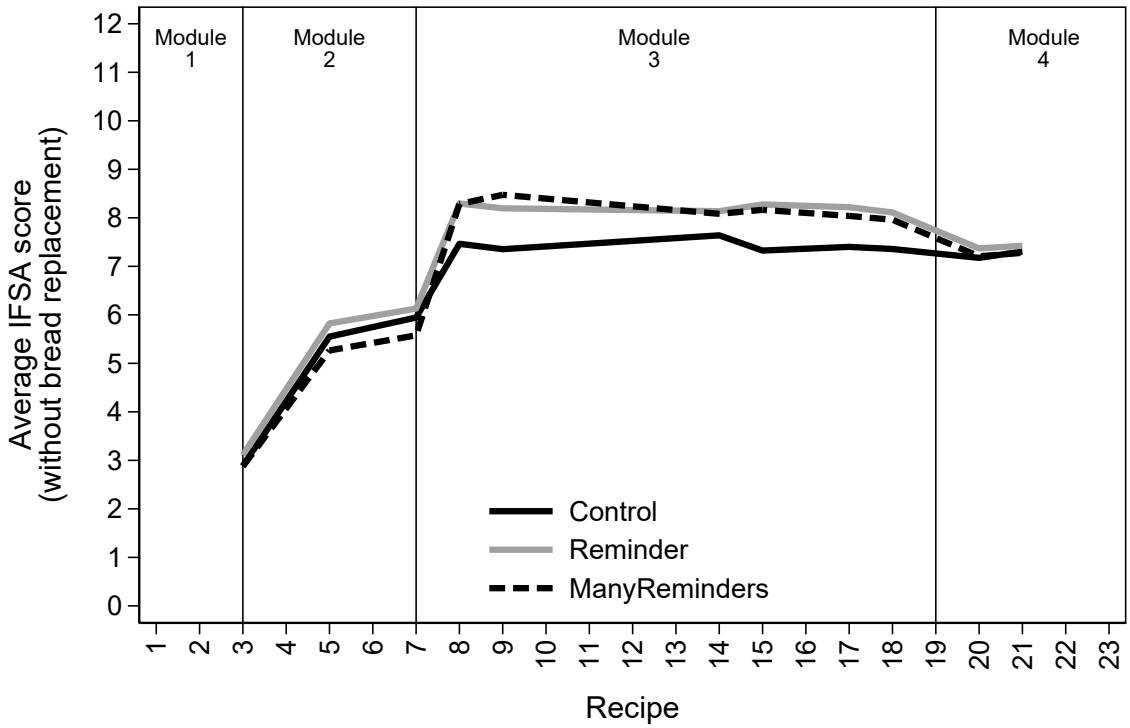
A.9 Additional figures

Figure A.1: Average IFSA score (Feedback study)



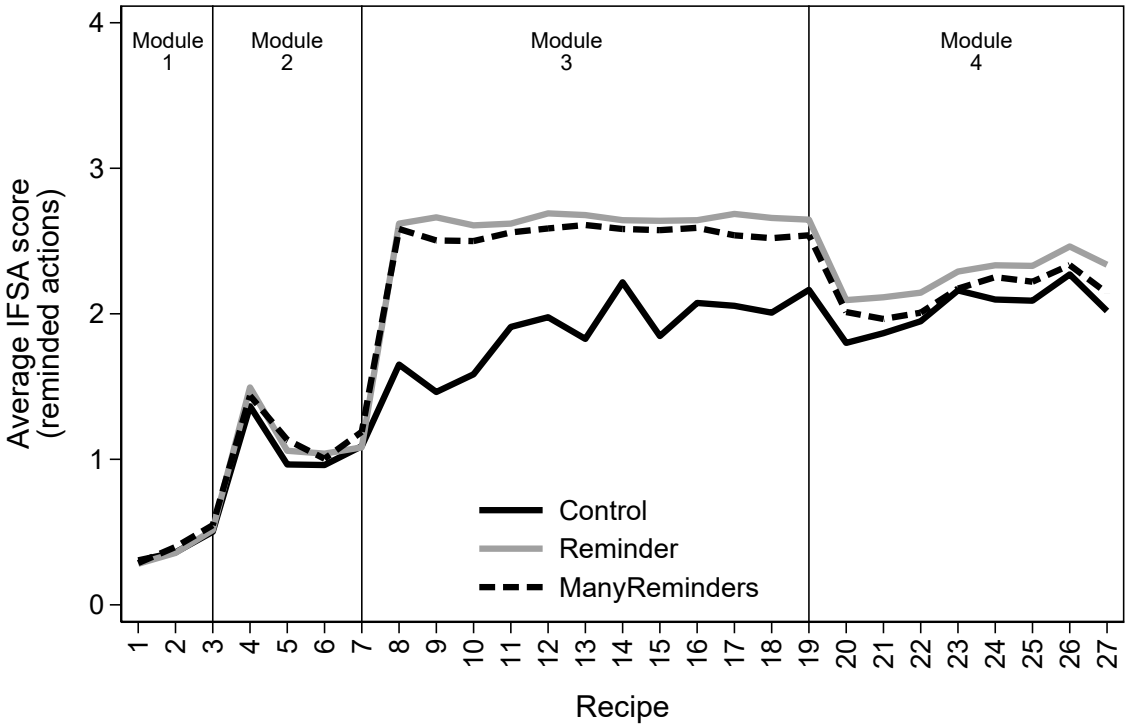
Notes: Mean of the IFSA score in the feedback study for each of the three treatments. Because bread only drops in some recipes, the score for throwing out dropped bread is excluded to obtain smoother graphs. Module 1 (recipes 1-3): Introduces the game mechanics. Module 2 (recipes 4-7): Starts with a survey, a video about safe food handling, and a video of a play-through of the game with correct food safety behavior. Then subjects start the recipes. Module 3 (recipes 8-19): Starts with telling subjects that their payment from now on will depend on how well they play the game: the bonus of up to GBP 3 is based on the IFSA score in one of the five game levels in modules 3 and 4, which each have four recipes. In *Reminder* (*ManyReminders*) subjects receive reminders about hand-washing (hand-washing, surface cleaning, and checking the temperature of the meat). After each level (four recipes), subjects receive feedback on their IFSA score for the level. Module 4 (recipes 20-27): Opens 48 hours after the first part of the study was posted on Prolific. Subjects are reminded that their payment depends on how well they play the game. The play the game without reminders. After each level (four recipes), subjects receive feedback on their IFSA score for the level.

Figure A.2: Average IFSA score (No feedback study)



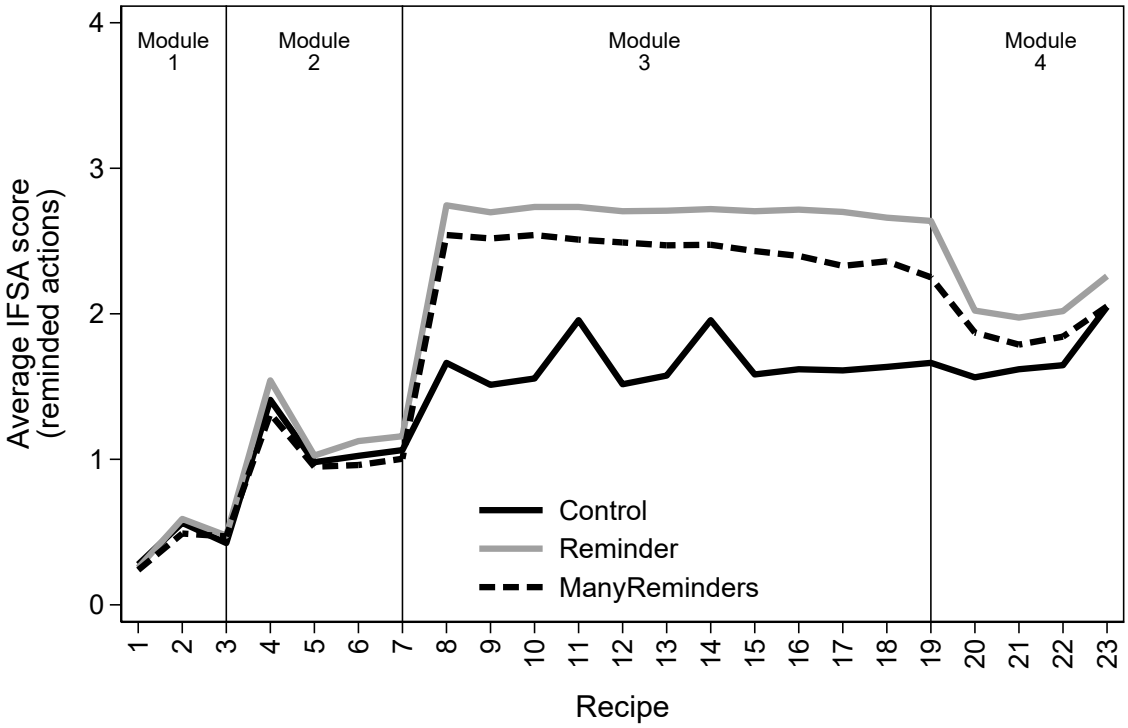
Notes: Mean of the IFSA score in the no feedback study for each of the three treatments. Because bread only drops in some recipes, the score for throwing out dropped bread is excluded to obtain smoother graphs. Subjects do not receive any feedback about their IFSA score in any part of the study. Module 1 (recipes 1-3): Introduces the game mechanics. Module 2 (recipes 4-7): Starts with a survey, a video about safe food handling, and a video of a play-through of the game with correct food safety behavior. Then subjects start the recipes. Module 3 (recipes 8-19): Starts with telling subjects that their payment from now on will depend on how well they play the game: the bonus of up to GBP 3 is based on the IFSA score in one of the five game levels in modules 3 and 4, which each have four recipes. In *Reminder* (*ManyReminders*) subjects receive reminders about hand-washing (hand-washing, surface cleaning, and checking the temperature of the meat). Module 4 (recipes 20-23): Opens 48 hours after the first part of the study was posted on Prolific. Subjects are reminded that their payment depends on how well they play the game. The play the game without reminders.

Figure A.3: Average IFSA score for the category of actions reminded in *Reminder* (feedback study)



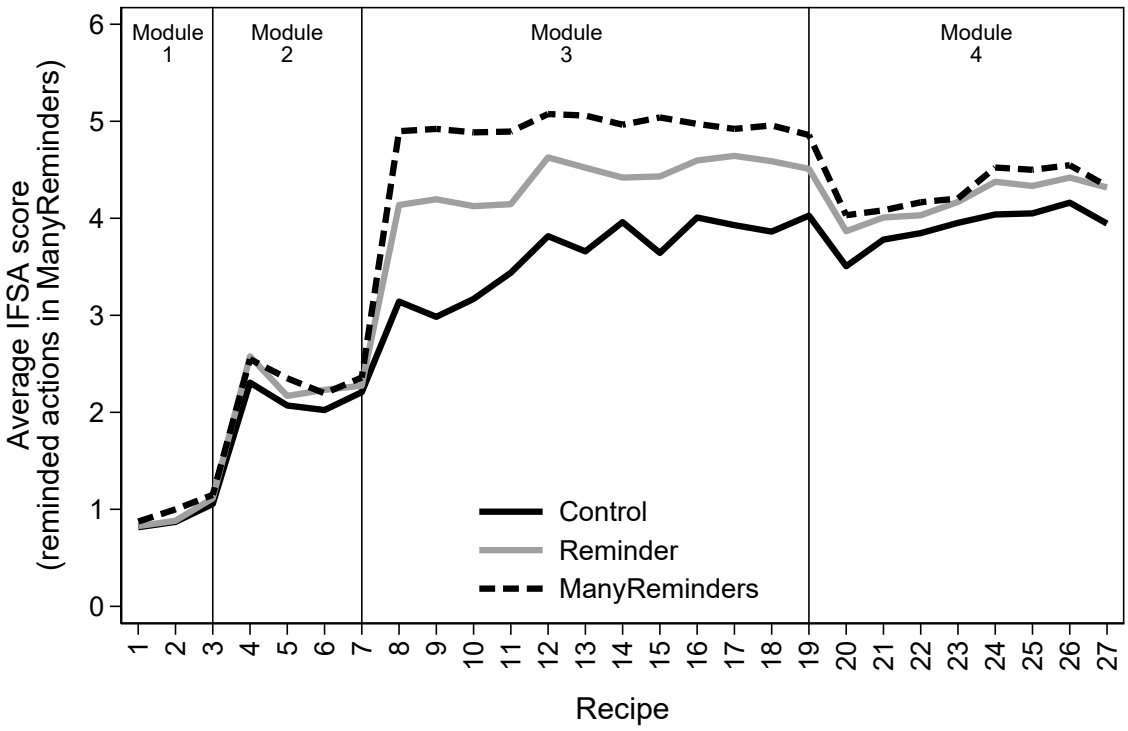
Notes: Mean of the IFSA score for hand-washing (the category of actions reminded in *Reminder*) in the feedback study for each of the three treatments. After recipes 11, 15, 19, 23, and 27 subjects receive feedback on their IFSA score for the level (four recipes). For further details, see the notes in Figure A.1.

Figure A.4: Average IFSA score for the category of actions reminded in *Reminder* (no feedback study)



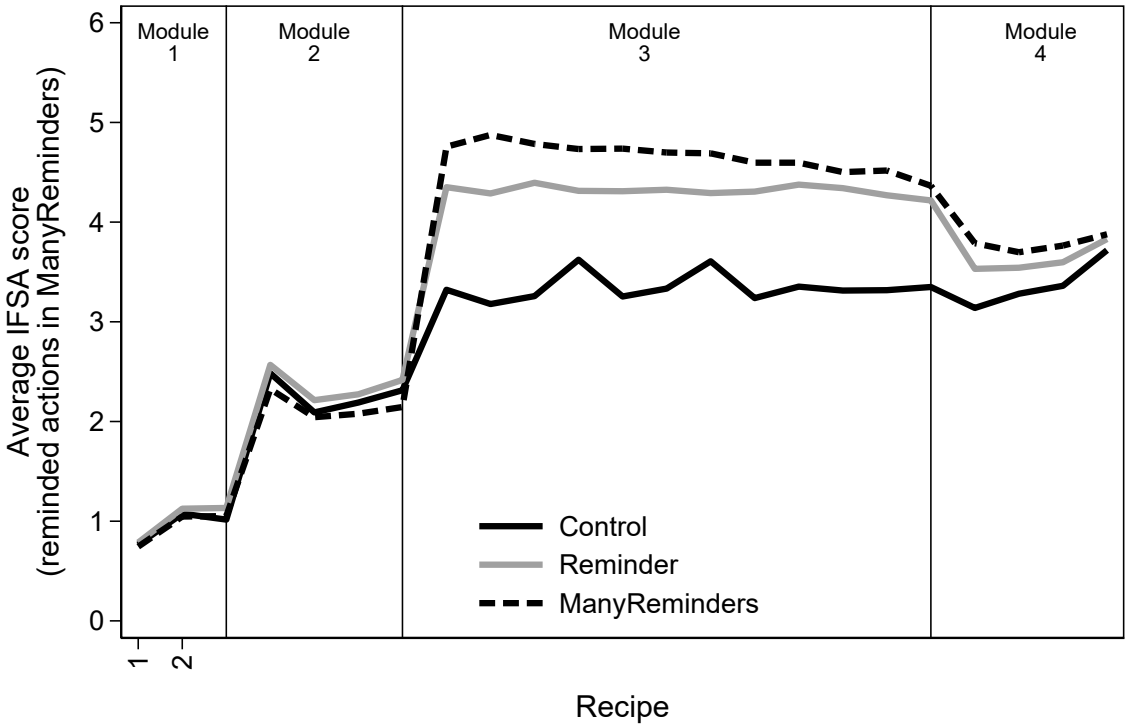
Notes: Mean of the IFSA score for hand-washing (the category of actions reminded in *Reminder*) in the no feedback study for each of the three treatments. Subjects do not receive any feedback about their IFSA score in any part of the study. For further details, see the notes in Figure A.2.

Figure A.5: Average IFSA score for the categories of actions reminded in *ManyReminders* (feedback study)



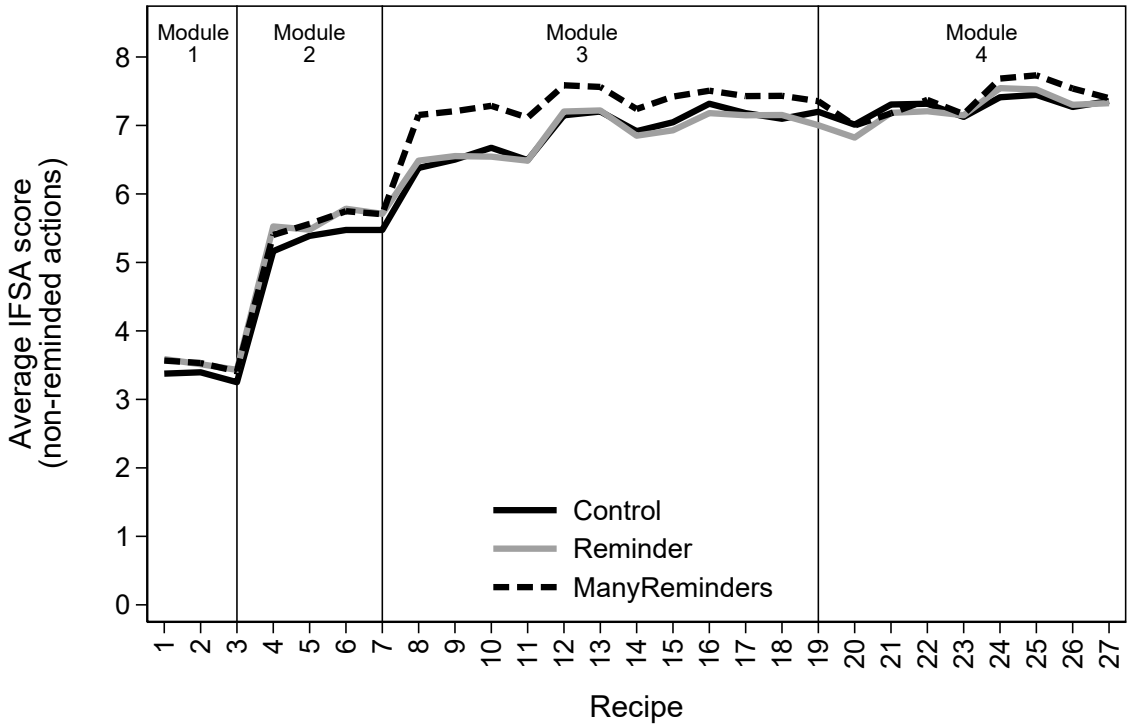
Notes: Mean of the IFSA score for hand-washing, surface cleaning, and checking the temperature of the meat (the categories of actions reminded in *ManyReminders*) in the feedback study for each of the three treatments. After recipes 11, 15, 19, 23, and 27 subjects receive feedback on their IFSA score for the level (four recipes). For further details, see the notes in Figure A.1.

Figure A.6: Average IFSA score for the categories of actions reminded in *ManyReminders* (no feedback study)



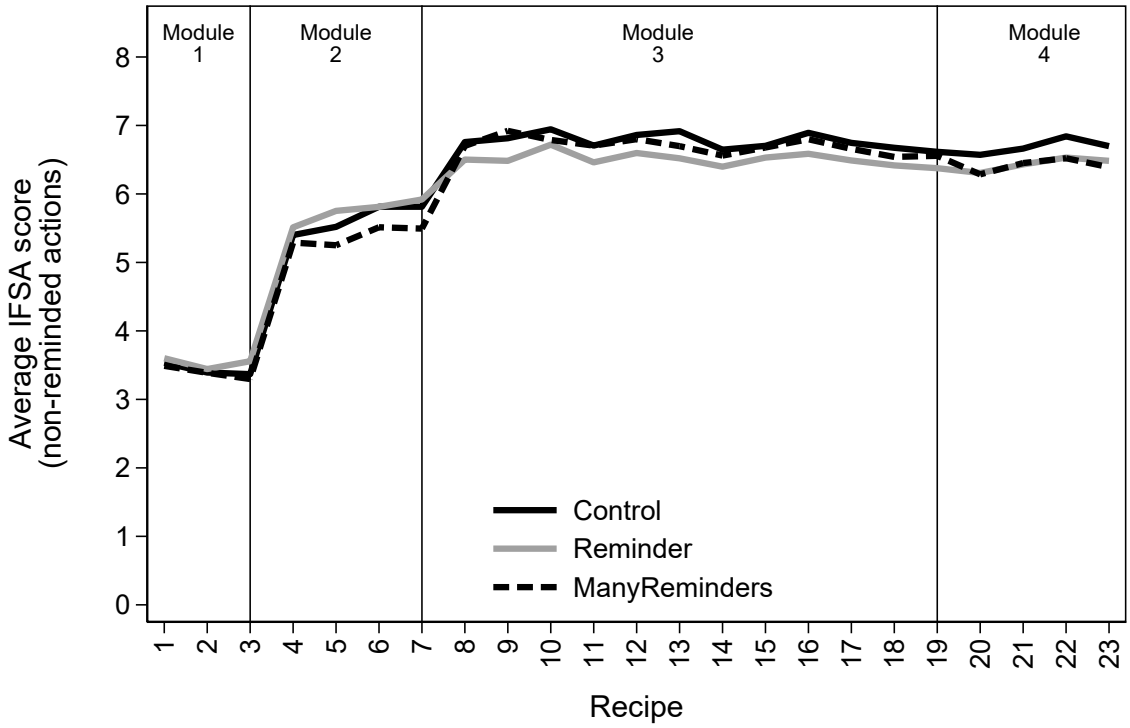
Notes: Mean of the IFSA score for hand-washing, surface cleaning, and checking the temperature of the meat (the categories of actions reminded in *ManyReminders*) in the feedback study for each of the three treatments. Subjects do not receive any feedback about their IFSA score in any part of the study. For further details, see the notes in Figure A.2.

Figure A.7: Average IFSA score for the categories of actions not reminded in *Reminder* (feedback study)



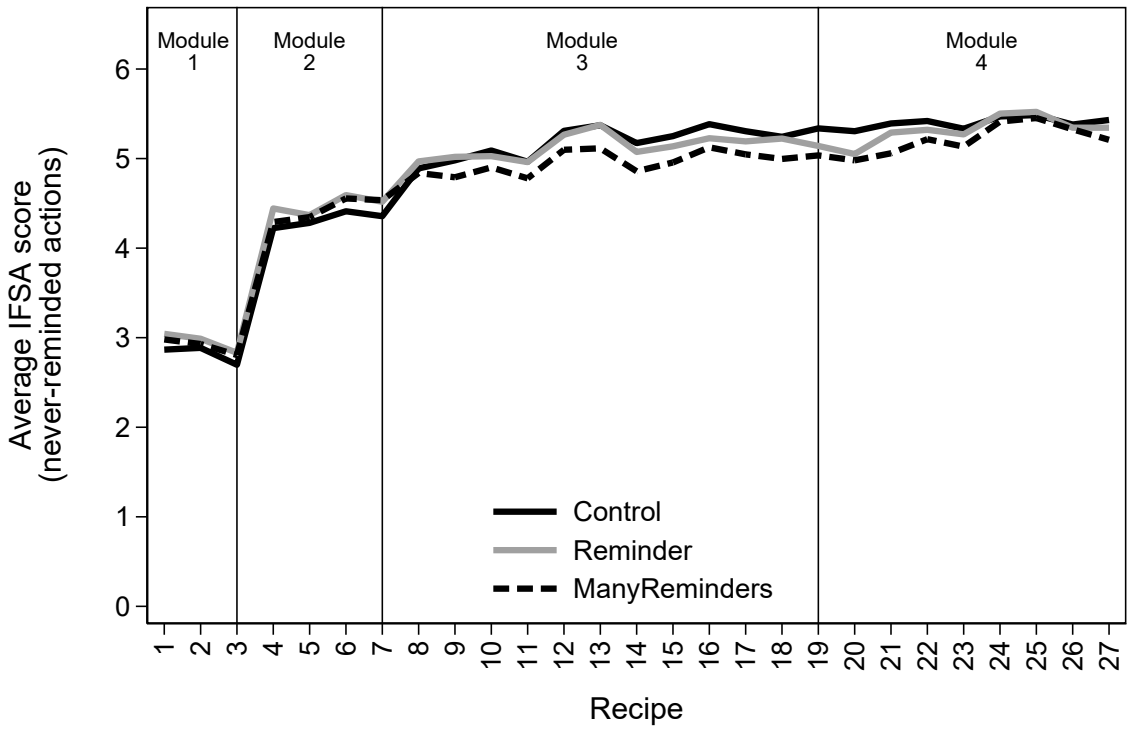
Notes: Mean of the IFSA score for all actions except hand-washing – the category of actions reminded in *Reminders* – in the feedback study for each of the three treatments. After recipes 11, 15, 19, 23, and 27 subjects receive feedback on their IFSA score for the level (four recipes). For further details, see the notes in Figure A.1.

Figure A.8: Average IFSA score for the categories of actions not reminded in *Reminder* (no feedback study)



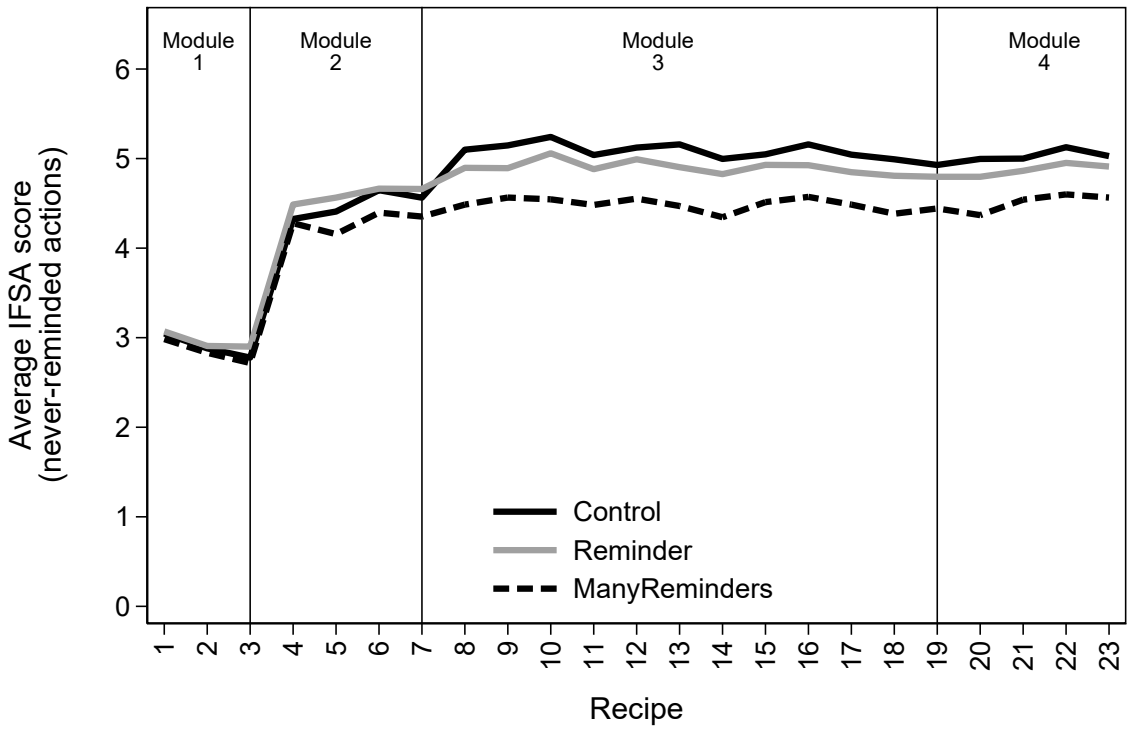
Notes: Mean of the IFSA score for all actions except hand-washing – the category of actions reminded in *Reminders* – in the no feedback study for each of the three treatments. Subjects do not receive any feedback about their IFSA score in any part of the study. For further details, see the notes in Figure A.2.

Figure A.9: Average IFSA score for the categories of actions not reminded in *ManyReminders* (feedback study)



Notes: Mean of the IFSA score for all actions except hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders* – in the feedback study for each of the three treatments. After recipes 11, 15, 19, 23, and 27 subjects receive feedback on their IFSA score for the level (four recipes). For further details, see the notes in Figure A.1.

Figure A.10: Average IFSA score for the categories of actions not reminded in *ManyReminders* (no feedback study)



Notes: Mean of the IFSA score for all actions except hand-washing, surface cleaning, and checking the temperature of the meat – the categories of actions reminded in *ManyReminders* – in the feedback study for each of the three treatments. Subjects do not receive any feedback about their IFSA score in any part of the study. For further details, see the notes in Figure A.2.