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

DISCUSSION PAPER SERIES

IZA DP No. 18273

Designing Effective Interventions

Sebastian Riedmiller
Matthias Sutter
Sebastian Tonke

NOVEMBER 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18273

Designing Effective Interventions

Sebastian Riedmiller
Max Planck Institute for Research on Collective Goods

Matthias Sutter
Max Planck Institute for Research on Collective Goods, University of Cologne, University of Innsbruck, IZA and CESifo

Sebastian Tonke
Max Planck Institute for Research on Collective Goods

NOVEMBER 2025

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

Designing Effective Interventions*

We provide a systematic framework to diagnose underlying problems and predict intervention effectiveness ex-ante. For this, we developed a parsimonious and generalizable survey tool (anamnesis). Our anamnesis classifies underlying problems along three fundamental diagnoses: awareness, intention, and implementation problems. We validate the framework in an online experiment with 7,500 subjects. We find that (i) intervention effectiveness is heterogeneous across different settings, and (ii) our diagnosis accurately predicts this heterogeneity. On average, predicting a 10%-effect corresponds to an actual effectiveness of 8.92%. We further demonstrate the applicability of our framework to predict heterogeneities in the setting of COVID booster take-up.

JEL Classification: C93, D01, D61, D90

Keywords: intervention design, heterogeneous treatment effects, context dependency, experiment

Corresponding author:
Matthias Sutter
Max Planck Institute for Research on Collective Goods
Kurt-Schumacher-Str. 10
53113 Bonn
Germany

E-mail: matthias.sutter@coll.mpg.de

* We thank John Beshears, Eric Bettinger, Elizabeth Linos, Stephan Meier, Katy Milkman, Todd Rogers, and Barnabas Szaszi for their helpful comments. Sebastian Riedmiller and Sebastian Tonke thankfully acknowledge support from the Joachim Herz Foundation. Parts of this paper were written while Sebastian Riedmiller was visiting Stanford University and while Sebastian Tonke was visiting Harvard University. Both thank these institutions for their hospitality. IRB approval was obtained from the University of Cologne (220006MS). This experiment was pre-registered at the AEA registry (AEARCTR-0009809). The Online Appendix is available at https://tinyurl.com/OnlineAppendixDEI.

1 Introduction

A core objective of economics is to design effective interventions. Unfortunately, many interventions fail to meet their target or show high context-specificity. White (2019) summarizes that around 80% of interventions across several domains show weak or no positive effects.[1] Other studies show that the effectiveness of interventions drops dramatically when scaled up or implemented in different settings (Allcott, 2015; Vivalt, 2020; DellaVigna & Linos, 2022; List, 2022). The conditions under which interventions work are barely identified (Bryan, Tipton & Yeager, 2021; Szaszi et al., 2022).

In this paper, we argue that the failure to design effective interventions is rooted in a lack of understanding of the underlying problem. Although it is clear that effective interventions hinge on understanding the underlying problem (Rodrik, 2010; Datta & Mullainathan, 2014; Rockenbach, Tonke & Weiss, 2025; Tonke, 2025), a systematic, generalizable framework to elicit and diagnose problems is missing. To fill this gap, we develop a framework to diagnose the type and to quantify the extent of the underlying problem, which in turn allows us to predict the effectiveness of specific interventions that are designed to address a particular problem type.

Our analytical approach is analogous to a medical consultation. Before an intervention is implemented, patients provide self-reported answers to a simple questionnaire (anamnesis). The anamnesis helps to identify the underlying problem (diagnosis). The diagnosis is used to predict the effectiveness of different intervention types and to recommend the most effective one (prescription). More generally speaking, policymakers typically design and implement interventions when a particular behavior does not match a stated (policy) goal.[2] This can occur, e.g., with respect to healthy nutrition, pro-environmental consumption, driving within the speed limit, saving for retirement, paying taxes, or labor productivity. An intervention then typically aims to narrow the gap between a policy goal and the behavior of individuals.

We argue that in order to design effective interventions, it is crucial to understand why people fail to act in line with a particular policy goal. We propose three fundamental reasons for such a failure. The first reason is an *awareness problem*: Subjects are unaware that their actual behavior deviates from their believed behavior. More precisely, their behavior is worse than they thought. The second reason is an *intention problem*: Individuals might not intend to meet the policy goal, thus failing to behave in line with the goal. The third reason is an

[1] In a similar spirit, a recent meta-study by Cala et al. (2025) finds that financial incentives are, on average, ineffective in increasing performance in the field. Meta-studies by Maier et al. (2022) and Mertens et al. (2022) argue that there is little evidence for the effectiveness of nudging after adjusting for publication bias.

[2] Note that such a policy goal may also be chosen by oneself, an employer, or some other entity that aims to change behavior.

implementation problem: Individuals fail to implement their intentions, even if they have the intention to meet the policy goal.

We develop a framework in which the prevalence of these three problems can be measured through a parsimonious and generalizable set of anamnesis questions. The answers from the anamnesis then result in a specific diagnosis, which can then be used to predict the effectiveness of different intervention types and make prescriptions for the most effective intervention.

To empirically validate the value of our framework, we show in an online experiment with 7,500 subjects that the same intervention type can succeed or fail to change behavior, depending on the underlying problem. To do so, we designed three experimental settings that exogenously induce an (i) awareness, (ii) intention, or (iii) implementation problem while holding other parameters constant. For each of these three settings, we implement three intervention types in addition to a baseline condition (without any intervention), resulting in a total of 12 experimental conditions. To resolve the awareness problem, we use reminders; to resolve the intention problem, we increase the monetary incentives; and to resolve implementation problems, we use simplifications. We measure task performance through a modified version of the real-effort task by Toussaert (2018). Participants have to remember 3-digit numbers appearing on their screen in short intervals. Upon request, they have to enter the last displayed number into an input field. Each participant receives 50 such queries. The stated policy goal is to answer all 50 queries correctly. This setup provides us with the necessary experimental control to induce an awareness, intention, and implementation problem separately, as well as high statistical power given our 12 experimental conditions.

We conduct our anamnesis among participants in each baseline condition of the three settings after participants have worked on the task. The anamnesis consists of only two questions, measuring intentions to meet the policy goal and beliefs about their performance. We elicit intentions by asking how many of the 50 queries the individual initially planned to answer correctly (0-50). We measure beliefs by asking how many of the 50 queries they think they answered correctly (0-50), incentivized by accuracy. Answering these two questions takes the average individual 36 seconds. The extent of an awareness problem is then measured by the difference between the beliefs about one's performance and actual performance. Intention problems are measured by the difference between one's stated intention to answer a certain number of queries correctly and the policy goal. Implementation problems are measured by the difference between one's stated intention and one's believed performance.

We find that the effectiveness of the three interventions strongly varies across settings. As pre-registered, we find that reminders are most effective when we induced an awareness problem, incentives are most effective in case of an intention problem, and simplifications are

most effective when an implementation problem was induced. This means that interventions are most effective when targeting the corresponding fundamental problem. Yet, interventions become much weaker in case of a mismatch of intervention and underlying problem. The key purpose of our framework is to generate point predictions about the treatment effects. We find that our predictions match actual intervention effects with high precision. Aggregating across settings and intervention types, we find that a predicted effectiveness of 10% ex-ante translates into an actual effectiveness of an intervention of 8.92%. We can also show that choosing an intervention based on our framework increases the treatment effect size by 58% compared to randomly choosing one of the three interventions.

Finally, we demonstrate the applicability of our framework in a field setting. For this purpose, we proceed in two steps. First, we show that the three fundamental problems can also be diagnosed using six qualitative questions (rather than quantitative ones) that ask participants on a 5-point Likert scale about their awareness, intention, and implementation problems. These qualitative questions may be less precise in measuring the extent of the underlying problem, yet they can be more practical to implement and easier to respond in some contexts. We find that these qualitative questions can predict the treatment effects in our online experiment very well. In a second step, we then use the validated qualitative questions and apply them to predict the outcomes of a large field experiment by Milkman et al. (2024). They ran a megastudy in 2022 using reminder messages and free Lyft rides to target take-up of COVID-19 booster vaccinations. Based on a new sample of 1,006 online participants, we use our framework to generate predictions about these interventions. In line with the actual treatment effects in their study, our diagnosis suggests that awareness problems are much larger than implementation problems. Hence, our framework predicts that reminders to get vaccinated should be much more effective than free Lyft rides, which is what Milkman et al. (2024) actually found.

Our study makes two main contributions. First, rigorous impact evaluations with the goal to find out "what works" have been conducted across a broad range of economic fields.[3] These rigorous impact evaluations have uncovered that the majority of interventions seem to fail when scaled up or implemented in different contexts (Allcott, 2015; White, 2019; Vivalt, 2020; DellaVigna & Linos, 2022; List, 2022). These findings call for novel approaches that can determine "what works when". Our paper fills this gap. We develop and test a simple framework that can predict the effectiveness of interventions across settings. Our anamnesis is easy to implement, generalizable, and succeeds in predicting which intervention types will

[3]E.g. development economics (Demeritt & Hoff, 2018), education economics (e.g. Jensen, 2010; Hoxby & Turner, 2015), behavioral finance (e.g. Duflo & Saez, 2003; Bhargava & Manoli, 2015), and environmental economics (e.g. Newell & Siikamäki, 2014; Bruelisauer et al., 2020).

be most effective in which setting.

We are unaware of other systematic and generalizable diagnosis tools to predict treatment effects of interventions, contingent on the type of underlying problem. There are, however, two alternative approaches to dealing with and predicting context dependency. The first approach is expert predictions. Yet, the existing results are not very encouraging in this respect, as, for example, individual-level predictions of laymen, professors, and practitioners poorly predict intervention effects in a given setting (DellaVigna & Pope, 2018a; Milkman et al., 2021b; DellaVigna & Linos, 2022; Milkman et al., 2024). The second approach is to hand-pick a set of interventions and to compare them empirically. This can be done simultaneously in so-called megastudies (Milkman et al., 2021a,b, 2024), by targeting interventions for a specific group based on their past behavior (Brody et al., 2023), through machine learning tools (Opitz et al., 2024), or in sequentially conducted adaptive experiments (Kasy & Sautmann, 2021). Testing many interventions is, however, often not possible for time, financial, and ethical constraints. Further, these approaches still require hand-picking a set of interventions ex-ante, potentially without systematic knowledge of the underlying fundamental problem. In fact, the lack of a systematic ex-ante diagnosis that allows for targeting specific problems with specific interventions may be the reason why the literature has documented so many failures of interventions that were intended to change human behavior (White, 2019; Bryan, Tipton & Yeager, 2021; Maier et al., 2022; Mertens et al., 2022; Szaszi et al., 2022; Cala et al., 2025).

Second, our paper relates to reviews of intervention types and taxonomies that link intervention types to underlying problems or models of behavior (Gneezy, Meier & Rey-Biel, 2011; Michie, Stralen & West, 2011; Datta & Mullainathan, 2014; Münscher, Vetter & Scheuerle, 2015; Benartzi et al., 2017; Szaszi et al., 2017; Engl & Sgaier, 2020; Löfgren & Nordblom, 2020).[4] In contrast to this literature, our primary goal is not to classify interventions, catalog them, and understand why they work. Our key contribution is to provide a practicable approach to diagnose the underlying fundamental problem within a systematic framework and to empirically show our framework's value in predicting an intervention's effectiveness, thus helping to choose the right intervention type. At the same time, the framework also addresses the question why so many interventions actually fail. They fail when there is a mismatch of the intervention with the underlying problem.

The paper is structured as follows. We first explain our framework and the prediction measures in Section 2. Then, we outline our experimental design and data used to validate our framework in Section 3. We present our hypotheses and results in Section 4. Section

[4]For a review on defaults, see Jachimowicz et al. (2019). For a review on commitment devices, see Bryan, Karlan & Nelson (2010).

5 introduces a qualitative version of our framework and provides evidence for its external validity. We conclude in Section 6.

2 The Framework

2.1 Diagnosis - Awareness, Intention, and Implementation Problems

Our diagnosis relies on a conceptual framework that categorizes discrepancies between actual performance and the policy goal into three types of fundamental problems, as shown in Table 1. The first fundamental problem is the awareness problem, which we define as being unaware that one's actual performance deviates from the believed performance. For example, individuals might be unaware that their current work performance is worse than they think, or they might be unaware that they are driving faster than they think. Such unawareness can stem from a lack of salience (Bordalo, Gennaioli & Shleifer, 2022), forgetfulness, or limited attention (Gabaix, 2019). If unawareness is the underlying fundamental problem, then the prescription would be an intervention that makes individuals aware of their deviation from their believed performance, for example, by giving them feedback about their own behavior or by sending reminders.

The second fundamental problem is the intention problem, which we define as the lack of intention to meet the policy goal after considering the available and perceived costs, benefits, and constraints. That is, whether and to which degree individuals intend to match a policy goal depends on the utility they derive from doing so. If matching the policy goal does not increase their utility, individuals do not form the intention to meet the policy goal. For example, individuals might not intend to match their employer's performance expectation, or might not want to drive within the speed limit or pay their taxes on time. If individuals have an intention problem, the prescription is to change the perceived costs, benefits, or constraints. For example, a policymaker might consider increasing the monetary payoff or correcting a common misperception about the costs of acting in line with a policy goal.

The third fundamental problem is the implementation problem. Individuals might fail to implement their intentions due to a lack of self-control (Thaler & Shefrin, 1981), procrastination (Laibson, 1997), unexpected complexity of the task, or other psychological factors (Ajzen, 1985, 1991). For example, task complexity might inhibit workers from implementing their intended productivity, or drivers might lack self-control to resist the temptation to speed. If individuals have implementation problems, policy interventions could aim to reduce the implementation costs through simplification, the removal of temptations, or commitment

devices.

2.2 Anamnesis - Identifying and Quantifying the Problem

Borrowing terminology from medical consultation, we call the process of coming to a diagnosis the anamnesis. Anamnesis describes the process of eliciting and analyzing individuals' self-reported behavior to diagnose problems. Two anamnesis questions suffice to make a diagnosis when the policymaker observes individuals' actual performance (α) and when the policy goal (γ) is known to the target population. The first question measures beliefs about one's performance (β) by asking, e.g., "How many tasks do you think you solved correctly?". The second question measures intention (i) by asking, e.g., "How many of the tasks did you initially intend to solve correctly?".

The awareness problem is measured as the difference between beliefs and actual performance (see the middle column in Table 1). For example, one might believe to have solved 40 tasks correctly while only 20 were actually solved. In that case, we diagnose an awareness problem of 20 tasks.

The intention problem is measured as the difference between the policy goal and the intended performance. For example, one might have intended to solve 40 tasks correctly, while the policy goal was to solve 50 tasks correctly. Here, we diagnose an intention problem of 10 tasks.

The implementation problem is measured as the difference between one's intention and beliefs about one's performance. Here, the individuals recognize that they failed to implement their original intention. For example, one might report having planned to solve 40 tasks but believes to have solved only 10 tasks. We diagnose an implementation problem of 30 tasks. One might wonder why the diagnosis does not rely on the difference between intention and the actual performance. The reason is that we want to measure awareness and implementation problems separately. We rely on the difference between intentions and beliefs to measure the implementation problem and the difference between the belief and actual performance to measure the awareness problem. Without an awareness problem, believed and actual performance are equal and can be used interchangeably.

2.3 From Diagnosis to Prescription

We now discuss how our diagnoses lead to predictions and prescriptions. First, we use our diagnoses to make predictions about the effectiveness of different interventions. Based on these predictions, we then recommend the intervention with the highest predicted effectiveness. This recommendation is called a prescription.

Table 1: The Framework: Definition and Anamnesis of Fundamental Problems

Diagnosis	Anamnesis	Prescription (Intervention)
Awareness problem Actual performance differs from believed performance.	**Belief (β) - action (α)**	• Reminders • Feedback about behavior • ...
Intention problem No intention to match policy goal after considering perceived costs, benefits, and constraints.	**Policy goal (γ) - intention (i)**	• Change incentives or costs • Correct misperceptions of costs and benefits (e.g. information) • ...
Implementation problem Failing to implement the intention.	**Intention (i) - belief (β)**	• Commitment devices • Reduction of implementation cost (simplification, planning prompts) • ...

Notes: The table shows the three fundamental problems of our framework. The anamnesis questions allow us to diagnose and quantify the extent of each problem. Typical interventions that are prescribed to solve the problems are provided in the last column.

We need to consider the following steps when making predictions. For now, and as pre-registered, we assume that our intervention only addresses *one single* problem. We use reminders for awareness problems, incentives for intention problems, and simplifications for implementation problems. Further, we assume that the intervention will resolve 100% of that specific problem. For example, after a reminder is implemented, we assume that there are no awareness problems anymore. Later, we will show that this is an overly optimistic assumption, but that our predictions are already quite precise nevertheless. They become even better once we empirically adjust these assumptions.

In addition, we assume that an intervention is ineffective for individuals who have "negative diagnosis values". For example, an individual who believes to perform even better than originally intended would have a negative score for the implementation problem. In such cases, there is no implementation problem, and hence, the respective prediction is set to zero.

The next important consideration is that individuals can have *multiple* problems simul-

taneously. For instance, someone might have both an intention and an implementation problem. In such cases, resolving the intention problem alone does not fully translate into behavioral change, as individuals still fail to implement a part of their intentions. As a consequence, only an adjusted share of the intention problem will cause mistakes that can be fixed through interventions. Below, we explain how our predictions can be adjusted to deal with concurrent problems.

Predicting Effectiveness of Interventions that Target Awareness Problems – The predicted effect (PE) of an intervention that addresses 100% of the awareness problem is equal to the diagnosed extent of the awareness problem, which is quantified as the gap between believed (β) and actual performance (α), as seen in Formula 1. An adjustment for concurrent problems is not necessary, as concurrent implementation and intention problems would reduce the believed performance and are hence already captured by lower values of β. The intervention will therefore fully resolve the awareness problem and translate into behavioral change.

$$PE(awareness) = max\{\beta - \alpha, 0\} \tag{1}$$

Predicting Effectiveness of Interventions that Target Intention Problems – To predict the effectiveness of an intention-tackling intervention, we need to consider whether there are concurrent awareness or implementation problems. If there are none, the gap between the policy goal and the intention can be fully closed by an intervention that addresses 100% of the intention problem. If individuals have concurrent awareness or implementation problems, however, an intervention that only addresses the intention problem will be less effective. Here, we have to take into account that, despite higher intentions, the individuals will still fail to implement a share of the tasks they intended to do. This could happen if there is a concurrent implementation problem and one's believed performance falls short of the initially intended performance. There can also be a concurrent awareness problem. Then, actual performance is smaller than the believed performance. As a result, we have a discrepancy between intention (ι) and actual behavior (α). For example, assume that someone performs only half as well as intended, i.e., manages to convert only 50% of the intended performance into actual performance due to concurrent problems. If an intervention now increases the intention and motivates an individual to attempt 10 additional tasks, we would argue that only 5 additional tasks will eventually be completed by the individual as the other 5 intended tasks are not converted into actual performance. Formula 2 shows how we adjust our prediction accordingly.

$$PE(intention) = \max\{\gamma - i, 0\} \cdot \left(1 - \frac{\max\{i - \alpha, 0\}}{i}\right) \tag{2}$$

The first factor shows the diagnosed extent of the intention problem, quantified by the gap between the policy goal (γ) and the intended performance (i). The second factor becomes relevant in the presence of concurrent problems. It calculates the share of the intended tasks that an individual manages to convert into actually completed tasks. If α equals i, for example, there are no concurrent problems, and subjects convert all of their intentions into actual performance. If α is smaller than i, there are concurrent problems as the actual performance is smaller than the intended performance. These concurrent problems decrease the second factor, meaning that the share of intention that is being converted into actual performance is also becoming smaller. Thus, the predicted effectiveness of an intention-addressing intervention decreases.

Predicting Effectiveness of Interventions that Target Implementation Problems - The extent of the implementation problem is quantified by the gap between the intended performance and the believed performance. Such an implementation problem exists if individuals know that they once intended to solve more tasks than they believe they actually did. To predict the effectiveness of an intervention that tackles 100% of the implementation problem, we need to consider whether there is a concurrent awareness problem. If there is none, the gap between the beliefs and the intention can be fully closed by an intervention addressing the implementation problem, as beliefs match actual performance in the absence of an awareness problem. If people have a concurrent awareness problem, however, the same intervention will be less effective. Beyond what the individual knowingly failed to implement, unawareness will cause the individual to still make mistakes. For example, assume that someone has an awareness problem and therefore actually only solves 80% of the tasks they believe they are solving. If an intervention now resolves the implementation problem and the individual believes they have solved 10 additional tasks, we would argue that, due to the concurrent awareness problem, only 8 of these additional tasks will actually be solved. The remaining 2 tasks will only be believed to be solved. Formula 3 shows how we adjust our prediction accordingly.

$$PE(implementation) = \max\{i - \beta, 0\} \cdot \left(1 - \frac{\max\{\beta - \alpha, 0\}}{\beta}\right) \tag{3}$$

The first factor measures the diagnosed extent of the implementation problem, which is quantified as the difference between intention (i) and the believed performance (β) on a task. The second factor accounts for potential awareness problems. It measures the share of

the implementation problem that is converted into actual performance. If α equals β, for example, there is no awareness problem, and resolving the implementation problem completely converts into actual performance. If α is smaller than β, there is a concurrent awareness problem, as actual performance is smaller than the believed performance. This decreases the second factor, meaning that unawareness causes a lower share of additional tasks to be solved by an intervention that targets the implementation problem. As a consequence, the overall predicted effectiveness of an implementation problem-addressing intervention decreases.

3 Experimental Test of the Framework

To test our analytical framework, we designed an online experiment with two purposes: First, to show that the effectiveness of an intervention hinges on the underlying fundamental problem, leading to heterogeneous treatment effects of the same intervention when applied to different problems. Second, to show that we can use our anamnesis to diagnose the fundamental problem and predict the effectiveness of interventions. With these goals in mind, we generated three settings in which we exogenously induced either awareness, intention, or implementation problems. Within each setting, we then tested three interventions (reminders, incentives, simplifications) against a baseline condition (without any intervention), leading to a total of 12 experimental conditions.

3.1 The Real-Effort Task

In our experiment, we used a modified version of the real-effort task (RET) by Toussaert (2018). Participants saw a three-digit number on their screen that changed every 1.2 seconds. In random intervals, they were asked to enter the last displayed number into an input field within 7 seconds. Example screens can be found in Appendix C. The task had an exception rule: If the last displayed number included the digit "3", participants had to enter only "0" into the input field to provide a correct answer. This exception was explained to participants in a salient manner in the experimental instructions. The task lasted for 10 minutes, during which participants were queried 50 times. The queries contained the digit "3" 20 times, which was explained to participants. The instructions stated that the goal is to answer all 50 queries correctly.

Participants' performance was incentivized using a loss framing. They received an endowment of \$2.68.[5] From all 50 queries, 5 were selected randomly to be payoff-relevant. For

[5]Participants were paid in £ as standard currency in Prolific. All values were multiplied by 1.14 based on the average exchange rate to US\$ during data collection from 22 Sep to 28 Nov 2022.

an incorrect answer, participants lost \$0.23, such that they could lose up to \$1.15 but end up with at least \$1.53. We chose a loss framing as it has been shown to effectively incentivize performance in both online settings (DellaVigna & Pope, 2018b) and the field (Hossain & List, 2012; Fryer et al., 2022). Participants could skip the task and watch a relaxing video instead. If they chose to do so, any remaining queries were counted as incorrect, thus decreasing their payment. The skip button was placed directly below the changing 3-digit number, which directed participants to the video.

After the task or the video, participants answered the anamnesis questions. In addition, we elicited sociodemographic information and measured economic preferences.[6] After the survey, participants were informed about their payments and were redirected to Prolific to finish the experiment. The experiment was programmed in oTree (Chen, Schonger & Wickens, 2016). The instructions for all experimental conditions can be found in the Online Appendix.[7]

3.2 Three Settings

Our objective was to generate three distinct settings that exogenously induce either an awareness, an intention, or an implementation problem. To achieve this, we modified the real-effort task described above to create three different settings (screenshots of the tasks are shown in Appendix C):

- **Setting AWA**: Lower salience of the exception rule.
  To induce awareness problems, we made the exception rule less salient in the instructions. Instead of being highlighted in its own paragraph, it was embedded within another paragraph alongside other important features of the real-effort task. We hypothesized that participants would be less likely to read the exception rule carefully and more likely to forget to apply it during the task.

- **Setting INT**: Flat fee for performance.
  To induce intention problems, we removed the piece-rate incentives. Participants received a flat fee of \$1.54 for participation and were explicitly informed that the number of correct answers would not affect their payment.

[6] We elicited age, gender, educational background, income level, occupational status, household size, size of resident city, and political preference. Following the preference survey module by Falk et al. (2023), we elicited risk, time, trust, altruism, and positive reciprocity preferences. Additionally, we included a question on competition preferences using the same format. The order of the preference questions was randomized.
[7] The Online Appendix is available at `https://tinyurl.com/OnlineAppendixDEI`.

- **Setting IMP**: Increased task complexity and temptation to skip RET.
  To induce implementation problems, we increased the implementation costs by making two adjustments. First, we increased the numbers displayed in the task from three to five digits, making it more tedious to memorize the last displayed number. Second, we made skipping the tasks more tempting by giving participants the option to skip the task entirely without having to watch a video for the remaining time.[8]

Interventions - We test three types of interventions – reminders, incentives, and simplifications – against a baseline of no intervention in each specific setting. The goal is to demonstrate that the effectiveness of the three interventions depends on the underlying problem and that our framework can predict the intervention's success. For each setting, we therefore have the baseline condition and the following three intervention groups, which results in a total of 12 experimental conditions:

- **Reminders**: Reminder of the exception rule.
  To address an awareness problem, we reminded participants to apply the exception rule. The reminder to enter "0" if the last displayed number contained a "3" was placed on the screen immediately before the task began and remained visible on top of the screen. This aimed to increase participants' awareness of the exception rule and prevent them from forgetting it during the task.

- **Incentives**: Increased pay for performance.
  To address an intention problem, we increased the monetary incentives. The initial endowment was set to \$3.25, and participants now lost \$0.34 for each incorrect answer among the five randomly chosen pay-off relevant queries. This aimed to increase the intended performance.

- **Simplifications**: Reduced task complexity and temptation to skip RET.
  To address an implementation problem, we simplified the task by requiring participants to memorize only two digits and eliminated the option to skip the task to remove any temptation. These modifications aimed to reduce the implementation costs and help participants answer as many queries correctly as intended.

3.3 Anamnesis

In our setting, we observe actual behavior, and the policy goal was communicated to participants in the instructions. Under these conditions, our anamnesis relies on two questions.

[8]The skip button immediately led to the anamnesis questions.

Answering these questions took participants on average only 36 seconds. To elicit intentions, we asked participants "Please be honest. Think back to the beginning of the study. How many of the 50 queries did you plan to answer correctly after reading the instructions?". Participants answered on a 0-50 scale. We elicited performance beliefs by asking "How many of the 50 queries do you think you answered correctly?" (0-50 scale). The belief elicitation was incentivized. Participants received \$0.11 for a correct guess, \$0.06 for a deviation of 1, and \$0.02 for a deviation of 2 queries. From the answers of participants in the baseline conditions, we diagnose the extent of the problems and make predictions about the effectiveness of the interventions, as explained in section 2.3.

Recall that we do not allow participants to have "negative problems", i.e., participants who are aware that they might have answered fewer queries correctly than they actually did are not considered to have an awareness problem. Similarly, someone who believes to have answered more queries correctly than intended is not considered to have an implementation problem. In these cases, the extent of the problem is set to 0. Stating an intention to solve more than 50 queries was impossible.

3.4 Sample, Randomization, and Balance

We collected data on the platform Prolific between September and November 2022. The sample consists of US participants fluent in English and with an approval rate of $\geq$95%. Participation from mobile devices was not allowed to maintain the functionality of the real-effort task. We use several measures, such as Captchas, honey pots, and attention checks, to prevent computer-generated answers in our experiment.[9]

Of the 8,312 participants who started the experiment, 106 were screened out or dropped out before being assigned to a treatment group, leaving 8,206 participants who were assigned to a specific experimental condition. Another 689 participants dropped out before finishing the experiment, and 17 failed the attention check, resulting in the pre-registered 7,500 participants who completed the experiment. We stratified treatment assignments by age, gender, and college education.[10] The experimental groups are balanced. The balance table for the intention-to-treat (ITT) sample (including drop-outs) is displayed in Table B.2.

[9]Participants were required to complete a Captcha. Those who failed were given a second attempt at a different Captcha. If they failed again, they were excluded from the study. Only one participant was excluded for failing both Captchas. Subsequently, participants had to enter their Prolific ID. We identified and excluded 64 cases of individuals attempting to participate multiple times. To further safeguard against non-human participants, a hidden honeypot question – requesting the participant's name but invisible to human respondents – was employed. As anticipated, no responses were recorded for this question, indicating that bot interference is negligible in our experiment. Finally, an attention check was embedded in the post-task survey. This check required participants to select a specific response on a Likert scale.

[10]We additionally balanced by gender through the option offered by Prolific.

Table B.1 shows the balance of the participants who finished the experiment. There is a slightly higher rate of dropouts in the implementation-problem groups. To show that this does not substantially confound our results, we will use our framework to predict average treatment effects for those who finished the experiment, as well as for the intent-to-treat sample in the Appendix. Figure A.1 displays a flowchart of the sampling process in detail. The average payment was $2.37 for 12:49 minutes, equivalent to $11.09 per hour.

4 Hypotheses and Main Results

This section describes the heterogeneous treatment effects across settings and whether our framework can predict them. The analysis is structured along our pre-registered hypotheses (AEARCTR-0009809).

4.1 Heterogeneous Treatment Effects Across Settings

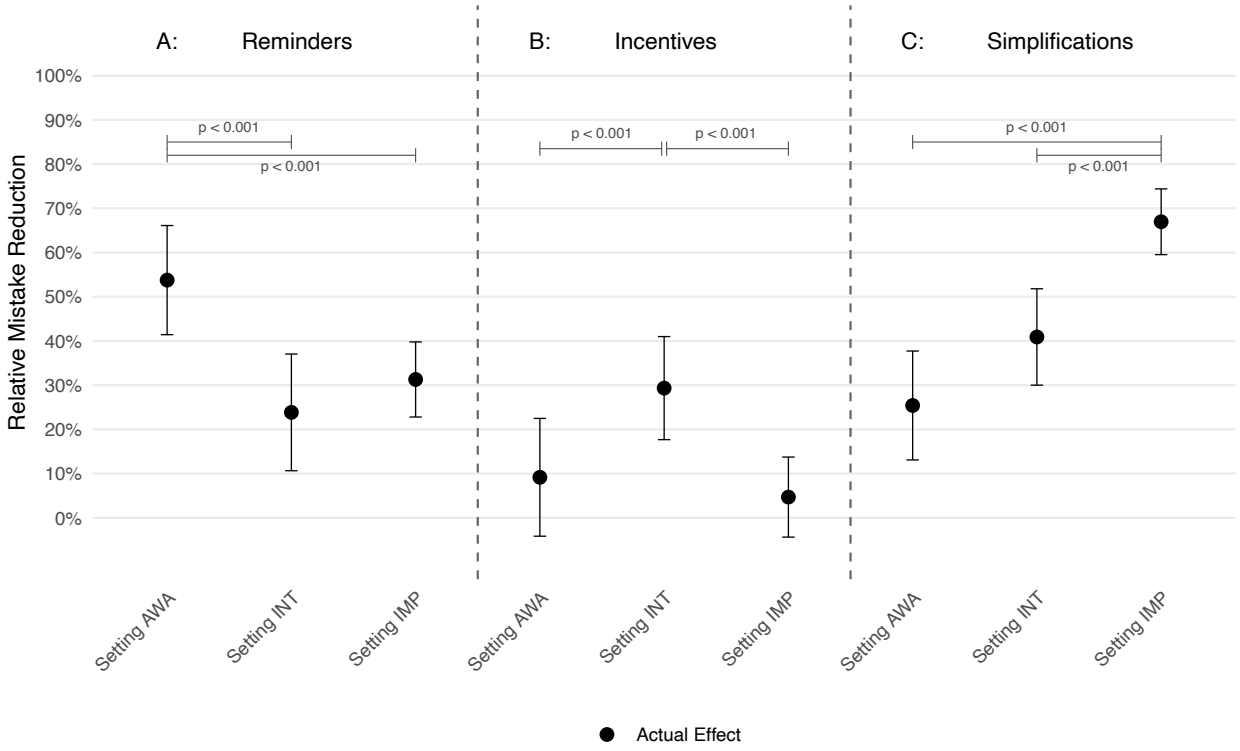
H1: *Reminders are most effective at reducing mistakes in Setting AWA.*

Panel A of Figure 1 displays the treatment effects of reminders in comparison to the baseline across settings. We normalize effect sizes by the respective baseline groups' average mistakes for comparability. That is, Panel A shows the percentage of baseline group mistakes that were resolved through the reminders. We find that the effectiveness of reminders varies substantially across settings. As hypothesized, reminders are most effective at reducing mistakes in Setting AWA. The reminders in the awareness setting reduce 54% of the baseline group's mistakes from 10.12 to 4.68, which is larger than in the intention and implementation setting (Wald-tests: $p<0.001$, $p<0.001$). Reminders are also effective in the other two settings, yet their effect size is significantly lower. In Setting INT, the actual treatment effect is 24%, reducing mistakes from 12.59 to 9.58. In Setting IMP, the actual treatment effect is 30%, reducing mistakes from 19.61 to 13.47. The corresponding regression statistics are shown in columns 1 and 2 of Table B.3.

H2: *Incentives are most effective at reducing mistakes in Setting INT.*

Panel B of Figure 1 shows the treatment effects of increased incentives across settings. As hypothesized, higher monetary incentives are most effective in Setting INT, where they reduce mistakes by 29% from 12.59 to 8.89. Treatment effects are significantly smaller in the other two settings (Wald-tests: $p<0.001$, $p<0.001$). The heterogeneity across settings is emphasized by the fact that we can not reject the null effects of incentives in Setting AWA and IMP. The corresponding regression statistics are shown in columns 3 and 4 of Table B.3.

Figure 1: Treatment Effect Heterogeneity Across Settings



Notes: This figure shows the intervention effects across the three settings. All effects are reported relative to the mistakes in the baseline groups per setting. The whiskers show the 95%-confidence intervals of the means. The displayed p-values refer to differences in treatment effects across settings.

H3: *Simplifications are most effective at reducing mistakes in Setting IMP.*

Panel C of Figure 1 shows treatment effects of the simplification across settings. In Setting IMP, they reduce mistakes by 67% from 19.61 to 6.48. As hypothesized, this effect is significantly larger than in the other settings (Wald-tests: $p<0.001$, $p<0.001$), where mistakes are reduced by 25% from 10.12 to 7.55 in Setting AWA, and by 41% from 12.59 to 7.44 in Setting INT. The corresponding regression statistics are shown in columns 5 and 6 of Table B.3 of the Appendix. We find similar patterns for the three hypotheses when using the ITT effects (Figure A.2 and Table B.4).

4.2 Predictions

We now evaluate whether our predictions are largest for the settings in which the actual treatment effects are strongest. Figure 2 plots predicted effect sizes on top of the actual treatment effects from the previous figure. As before, the predicted effect sizes are normalized

by mistakes in the respective baseline group. Recall that we assume here that reminders resolve the awareness problem entirely, that incentives resolve the intention problem entirely, and that the simplifications resolve the implementation problem entirely (see section 2.3). Later, in Section 4.3, we will refine these assumptions.

H4: *A higher predicted effectiveness of reminders corresponds to higher actual effectiveness.*

Overall, we find strong empirical support in line with H4. As shown in Panel A of Figure 2, we predict that reminders resolve 48% of the mistakes in the awareness problem setting, which is close to the actual effect size of 54%. As hypothesized, the predicted effect in Setting AWA is significantly larger than in the other two settings (t-tests: $p<0.001$, $p<0.001$). In Setting INT, we predict that reminders could reduce mistakes by 30%, while the actual treatment effect is 24%. In Setting IMP, we predict an effectiveness of 19%, while the actual treatment effect is 30%. Note, that while we did not intend to induce awareness problems in Settings INT and IMP, our framework nevertheless accurately predicts reminders to be effective in those settings, and to a smaller extent than in Setting AWA.

H5: *A higher predicted effectiveness of incentives corresponds to higher actual effectiveness.*[11]

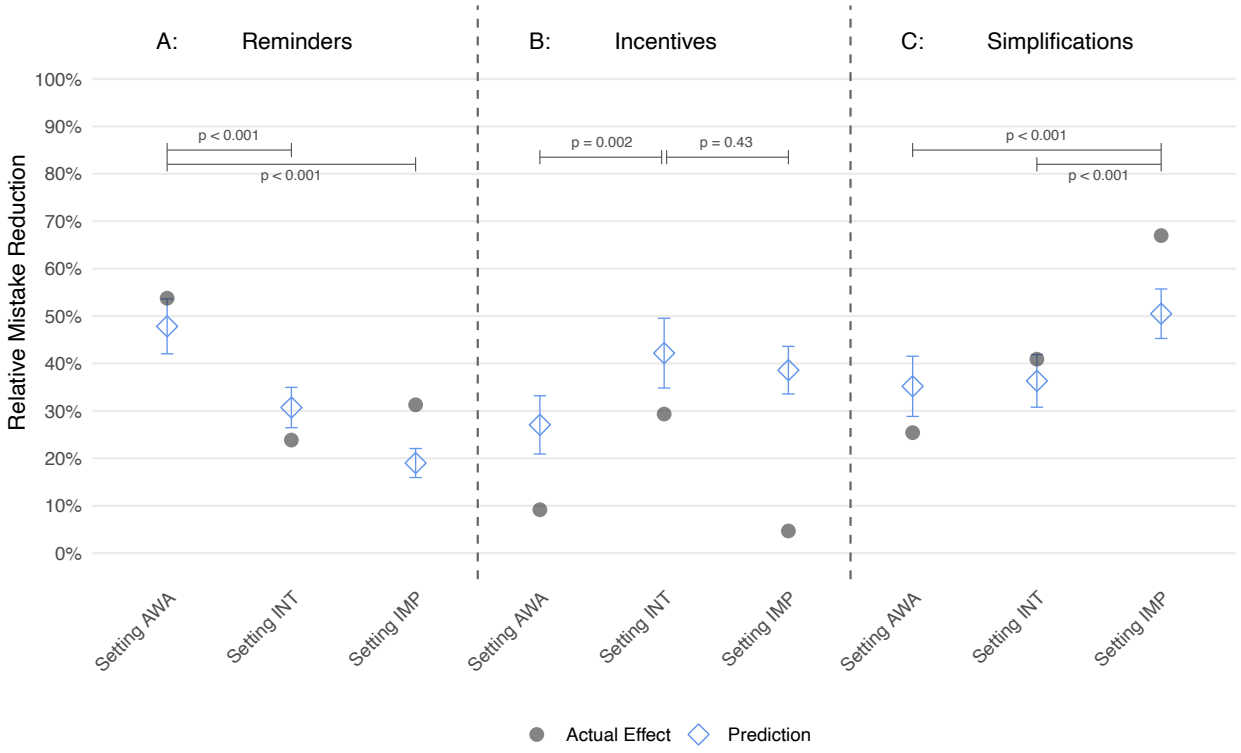
Panel B of Figure 2 shows the predicted treatment effects for increased incentives. In line with actual treatment effects, predictions are the largest for Setting INT, with a predicted reduction of mistakes of 42%. This is significantly larger than the predicted effect for Setting AWA (t-test: $p=0.002$), but not Setting IMP (t-test: $p=0.43$). While we find partial support for H5, the predictions seem to overshoot. We predict an effectiveness of increased incentives in Setting AWA of 27% and of 39% in Setting IMP, whereas actual treatment effects are not significantly different from zero. An explanation for this overshooting is that the incentives were not large enough to address intention problems, which we discuss in section 4.3.

H6: *A higher predicted effectiveness of simplifications corresponds to higher actual effectiveness.*

Panel C of Figure 2 shows the predicted effect of the simplifications. As hypothesized and in line with the actual treatment effects, we find that predictions are the largest for the implementation problem setting. In Setting IMP, we predict a reduction of 51%, which is significantly larger compared to the other two settings (t-tests: $p<0.001$, $p<0.001$), where we predict reductions of 35% and 36%.

[11] The predicted effectiveness is the share of intention-problem-driven mistakes that can be reduced by resolving the intention problem alone (see Section 2.3). In other words, only a share of the intention problem will cause mistakes that can be fixed through interventions due to concurrent problems. In our preregistration, we wrote that "a higher share of diagnosed intention barriers predicts higher effectiveness", which is less precise.

Figure 2: Predictions and Actual Treatment Effects



Notes: This figure shows the predicted and actual intervention effects across the three settings. The predicted effects are calculated as described in section 2. All effects are reported relative to the mistakes in the baseline groups per setting. The whiskers show the 95%-confidence intervals of the means. The displayed p-values refer to differences in predicted effects across settings.

Beyond understanding whether the ordering of the predictions matches actual treatment effects, one might also wonder how well the size of the treatment effect is predicted. To do so, we use OLS regressions in Table 2. Since we do not observe treatment effects on the individual level, we regress the average predictions on the average treatment effects and use bootstrapping to obtain the standard errors. Panel A shows that for reminders, a 1 percentage point reduction in predicted effectiveness translates to a 0.851 percentage point reduction ($p<0.001$) in actual effectiveness. For incentives, we find a high correlation coefficient of 0.912, yet this coefficient is not statistically significantly different from zero. For simplifications, we find a coefficient of 2.345 percentage points ($p<0.001$). When pooling all intervention types and predictions in Column 4, we find a coefficient of 1.196 ($p<0.001$). Since we normalized baseline mistakes, this can be interpreted as a 10% prediction resulting in a 11.96% reduction of mistakes. These findings provide strong evidence that our framework can predict the treatment effects.

Next, one might be interested in how effective our predictions are from an ex-ante per-

Table 2: Prediction of Intervention Effects by the Framework

	Actual Mistake Reduction			
	Reminders (1)	Incentives (2)	Simplification (3)	Pooled (4)
Panel A:				
Predicted Mistake Reduction	0.851***	0.912	2.345***	1.196***
	(0.211)	(0.740)	(0.664)	(0.198)
Panel B: No Intercept				
Predicted Mistake Reduction	1.085***	0.416***	1.128***	0.892***
	(0.082)	(0.079)	(0.062)	(0.045)

Notes: This table shows OLS results of the predicted mistake reduction in percent on the actual mistake reduction due in percent. We use the diagnosis of the baseline settings in each of the three settings to predict treatment effects. To obtain standard errors, we use bootstrapping to resample the original sample 1,000 times, calculate the mean intervention and predicted effects for each setting, and perform the OLS analysis on the aggregate data with and without allowing for an intercept. The standard deviations of these bootstrapped coefficients are used as standard errors and reported in parentheses. ***p<0.01

spective, i.e., when the intercept with the y-axis is unknown. Panel B of Table 2 regresses the average predictions on the average treatment effect but without estimating an intercept (i.e., forcing the prediction of zero to go through the null point). For reminders, we find that a predicted mistake reduction of 1 percentage point translates into an actual mistake reduction of 1.085 percentage points, which is remarkably accurate. For incentives, we find a correlation coefficient of 0.416, which corroborates that our predictions overshoot the actual effectiveness of the incentives. For simplifications, we find a coefficient of 1.128. Hence, our predictions for the simplification are close to 1, yet slightly undershoot the actual effects. When aggregating predictions across all settings, our regression suggests that a 10% prediction translates into an actual effectiveness of 8.92%, which underlines the accuracy of our prediction. We find similar results using predictions when ignoring concurrent problems and when using the ITT sample, displayed in Tables B.5 and B.6 in the Appendix. In sum, our results demonstrate that treatment effects are heterogeneous across settings, and that our framework can predict this.

What are the benefits of using our diagnostics compared to randomly selecting an intervention? The average effect of the interventions with the highest predicted effectiveness per setting is 50.02% while the average effect of all interventions in our settings is 31.71%. That means that prescribing an intervention based on our framework increases the effect size by 58% compared to randomly choosing one of the tested interventions, which highlights the value of conducting a diagnosis before choosing an intervention.

4.3 Improving Predictability by Refining the Assumptions

While our predictions are quite precise overall, they perform better for some interventions than others. Most notably, we find that the predictions for the monetary incentives were too high compared to the actual treatment effects. An explanation for this is that our assumptions regarding the effectiveness of the interventions were too optimistic. In other words, the additional monetary incentive may not have resolved 100% of the intention problem. Our data allows us to analyze the degree to which an intervention affected the underlying problem by comparing differences in diagnoses between the baseline and the treatment groups, which is shown in Table 3. If our assumptions about the effectiveness of our interventions were true (see Section 2.3), we would expect that the reminders completely solve the awareness problem but not the intention and implementation problems. That is, in Panel A of Table 3, column 1 would show values of -1, and columns 2 and 3 would show 0. Equivalently, for Panel B, we would expect values of -1 in column 2 and 0 otherwise. For Panel C, we would expect values of -1 in column 3 and 0 otherwise.

Column 1 of Panel A shows the extent to which the reminders actually reduce the awareness problem. In Setting AWA, reminders reduce the diagnosed awareness problems by 81%. In Settings INT and IMP, where the awareness problem was smaller to begin with, reminders reduce the awareness problem by 56% and 53%. Columns 2 and 3 show that reminders did not significantly affect the diagnoses of intention and implementation problems. We conclude that our predictions for reminders are quite precise because our reminders largely work as assumed: They consistently reduce awareness problems without impacting the other two problems.

The assumptions regarding the incentives were too optimistic. As shown in Panel B, column 2, the incentives in Setting INT solved half of the intention problem (53%), reducing its diagnosed extent from 6.33 to 2.95 mistakes. Interestingly, we find that the monetary incentive slightly increased awareness problems by 22% (column 1) and reduced implementation problems by 31% (column 3). In the other two settings, the intention problem was not resolved at all, and neither of the other problems was affected. Hence, substantial intention problems remain in those two settings, as the extra monetary incentive may have been insufficient to tackle the intention problem. Why were incentives in setting INT more effective in addressing the intention problem? In Setting INT, the relative increase in payments was larger. Participants in the baseline received no pay for performance, whereas the treatment group received a potential bonus of $1.71, compared to the other two settings where the bonus was just increased from a possible $1.14 to $1.71. The marginal increase of the monetary incentive in the other two settings seems to have been too small to resolve

Table 3: Relative Intervention Effects on Diagnosed Problems

	Change in Diagnosed Problems		
Experimental Group	Awareness (1)	Intention (2)	Implementation (3)
Panel A: Reminders			
Setting AWA	-0.81***	-0.20	0.00
Setting INT	-0.56***	-0.07	0.03
Setting IMP	-0.53***	-0.09	-0.12
Panel B: Incentives			
Setting AWA	-0.08	0.08	-0.12
Setting INT	0.22**	-0.53***	-0.31**
Setting IMP	-0.03	-0.03	-0.07
Panel C: Simplification			
Setting AWA	0.05	-0.49***	-0.42***
Setting INT	0.01	-0.62***	-0.32***
Setting IMP	0.06	-0.79***	-0.77***

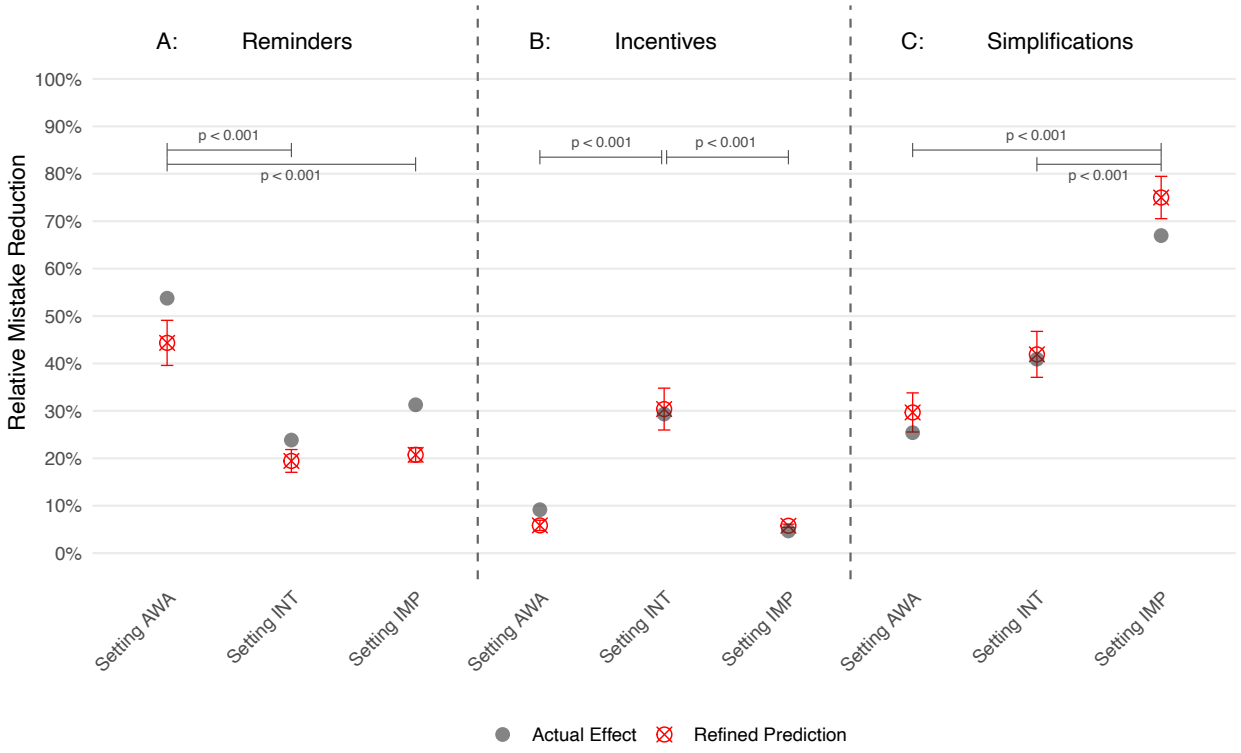
Notes: The table shows the intervention effects on the extent of diagnosed problems for each experimental setting. All values show the relative change of diagnosed problems relative to the baseline group of the respective setting. Negative values indicate that the extent of the underlying problem was reduced.

the intention problem. Since our intervention was not as effective in tackling the underlying problem, our predictions in Figure 2 overshoot.

Panel C of Table 3 evaluates our assumption regarding simplifications. As intended, simplifications reduced implementation problems across all settings (column 3). For example, in Setting IMP, simplifications solved 77% of the implementation problem, reducing its diagnosed extent from 11.09 to 2.53 mistakes. Awareness problems are not affected by simplifications (column 1). We did not anticipate, however, the degree to which the simplification also addresses intention problems. Column 2 shows that simplifications substantially reduced intention problems between 49% and 79%. As a result, actual treatment effects overshoot their predictions as simplifications not only resolve implementation problems but also intention problems.

Could our predictions be improved if we had a better knowledge of the extent to which treatments address the underlying problems? To see whether refined assumptions also lead to better predictions, we use the values from Table 3 instead of assuming that an intervention addresses only one type of problem and resolves it by 100%. Understanding whether refined assumptions improve predictability is important for two reasons. First, if the changes in diagnoses between the intervention and baseline groups contain meaningful information, we would expect more accurate predictions. This would underscore the validity of our framework. Second, policymakers and researchers might learn to form (or already have)

Figure 3: Precision of the refined predicted intervention effect across settings



Notes: This figure shows the refined prediction of intervention effects and actual intervention effects across the three settings. The refined predictions are calculated as described in section 4.3. All effects are reported relative to the mistakes in the baseline groups per setting. The whiskers show the 95%-confidence intervals of the means. The displayed p-values refer to differences in the refined predicted effects across settings.

more accurate assumptions about how treatments affect the underlying problem types.

To make the refined adjustments, we simply multiply our predicted effects (Section 2.3) by the fraction of how well an intervention actually resolves the respective fundamental problem in each setting. Furthermore, we relax the assumption that an intervention may only affect one of the underlying problems. Here, the predictions for affecting the awareness, intention, and implementation problems are added per intervention and discounted by the degree to which they actually reduce the underlying problem.[12]

Figure 3 plots the actual treatment effects and the refined predictions. The predictions now come even closer to the actual treatment effects. Table 4 provides corresponding regression statistics (all p-values<0.001). When pooling across all treatments (column 4), we find that a 10% increase in the refined predicted effect translates into an 8.83% increase in the actual effect (Panel A), in comparison to 11.96% without refinement (Table 2). For

[12] $$RefinedPE_i = \max\{\beta-\alpha,0\}\cdot\delta^i_{aw} + \max\{\gamma-i,0\}\cdot\delta^i_{in}\cdot\left(1-\frac{\max\{\beta-\alpha,0\}\cdot\delta^i_{aw}+\max\{i-\beta,0\}\cdot\delta^i_{im}}{i}\right) + \max\{i-\beta,0\}\cdot\delta^i_{im}\cdot\left(1-\frac{\max\{\beta-\alpha,0\}\cdot\delta^i_{aw}}{\beta}\right)$$

Table 4: Refined Prediction of Intervention Effects by the Framework

	Actual Mistake Reduction			
	Reminders (1)	Incentives (2)	Simplification (3)	Pooled (4)
Panel A: Aggregate Level				
Refined Predicted Mistake Reduction	1.090***	0.913***	0.890***	0.883***
	(0.267)	(0.266)	(0.122)	(0.055)
Panel B: Aggregate Level - No Intercept				
Refined Predicted Mistake Reduction	1.260***	0.981***	0.907***	0.994***
	(0.104)	(0.170)	(0.042)	(0.051)

Notes: This table shows OLS results of the refined predicted mistake reduction. The predicted and the actual mistake reduction are standardized by the average mistakes of the baseline groups in each setting. Regression coefficients are in percentage terms. Standard errors are bootstrapped by resampling the original sample 1,000 times. We calculate the mean intervention and refined predicted effects for each setting, and perform the OLS analysis on the aggregate data with and without allowing for an intercept. The standard deviations of these bootstrapped coefficients are used as standard errors and reported in parentheses.
***p<0.01

the ex-ante prediction, we even find that a 10% predicted effectiveness translates into an actual effectiveness of 9.94% (Panel B), in comparison to 8.92% without refinement (Table 2). These results highlight the potential of our framework to predict treatment effects with high accuracy.

5 Qualitative Anamnesis and External Validity

So far, we have demonstrated that our framework can predict treatment effects in a controlled online setting. This online setting was ideal to exogenously induce the three problem types and test the effectiveness of the three interventions among a high-powered sample. In settings with less control, it might be useful to use qualitative questions as anamnesis. The qualitative questions are less precise in measuring the extent of the underlying problem, yet they might be more easy to apply in certain settings. In a first step, we will demonstrate that a qualitative anamnesis has a similar predictive power as the previously introduced quantitative anamnesis in our online experiment. Based on this finding we then examine in a second step the external validity of the qualitative anamnesis. We do so by providing evidence that our framework can predict treatment effects from a field experiment by (Milkman et al., 2024).

5.1 Qualitative Anamnesis and Diagnosis

Qualitative Anamnesis Questions - We administered the qualitative anamnesis questions to participants who answered at least one query incorrectly. For each problem type, we asked two questions that can be answered on a 5-point Likert scale ranging from "Definitely Yes" to "Definitely No". Answering these questions took participants on average 62 seconds.

To measure awareness problems, we asked: "Are you surprised that you answered exactly X queries incorrectly?" and "Did you forget to apply the extra rule at any point during the task?"[13]. To measure intention problems, we asked: "Think back to the beginning of this study: Did you ever plan to answer all 50 queries correctly after reading the instructions?" and "Think back to the beginning of this study: Were you ever determined to answer all 50 queries correctly after reading the instructions?" To measure implementation problems, we asked: "Did you consciously decide to answer fewer queries correctly than you planned at the beginning of this study? Note: All skipped and unanswered queries are counted as incorrect." and "Did you have difficulties answering as many queries correctly as you planned at the beginning of this study?"

Diagnosis and Concurrent Problems - We diagnose a problem if both anamnesis questions are answered with "Definitely Yes (No)" or "Rather Yes (No)".[14] As previously, we take into account concurrent problems. For the qualitative anamnesis, we assume a simple rule to predict intervention effectiveness: If we diagnose one problem type only, 100% of the gap between behavior and the policy goal is resolved by an intervention that tackles this specific problem. If we diagnose two concurrent problems, an intervention targeting only one problem type will only resolve half of the gap. With three concurrent problems, an intervention that targets one specific problem will only resolve a third of the gap.

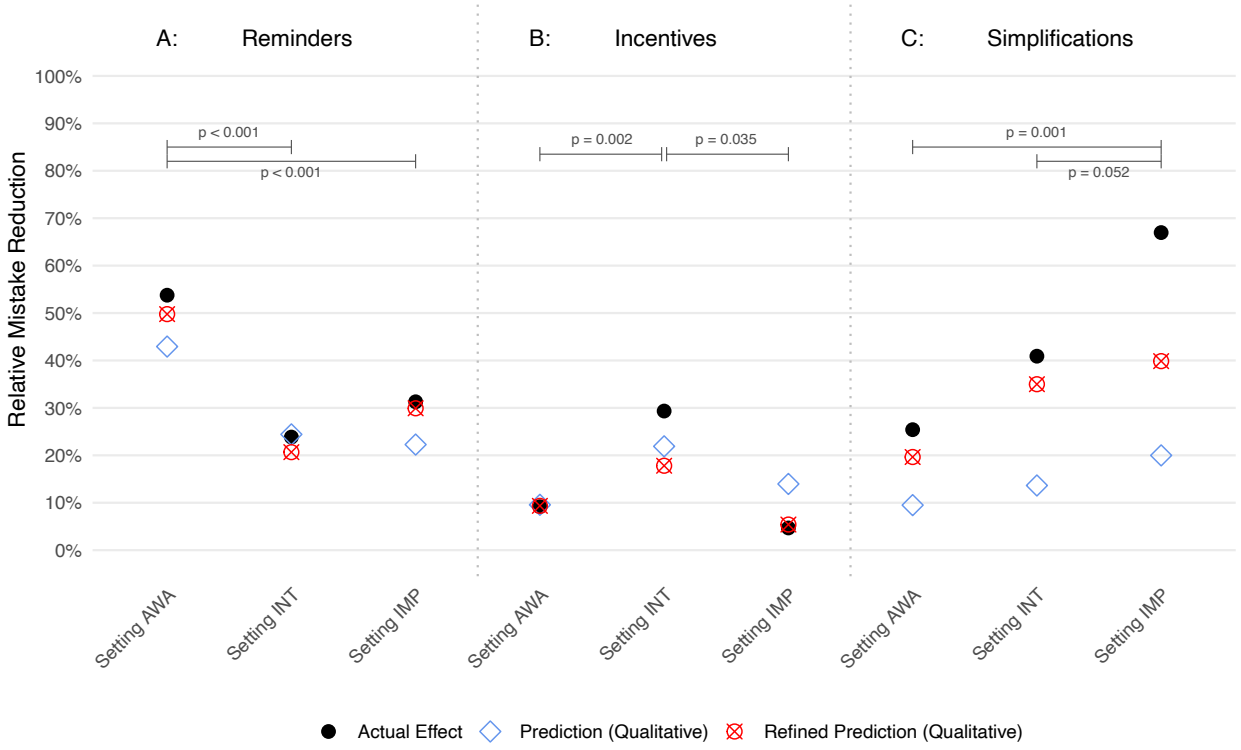
Refined Predictions - For the refined predictions, we use the same approach as for the quantitative anamnesis (Table 3). We analyze the degree to which the interventions affected the underlying problem by comparing differences in diagnoses between the baseline and the treatment groups, which is shown in Table B.7.

As before, we calculate the refined prediction by allowing an intervention to affect all three problem types, e.g., simplifications can affect awareness, intention, and implementation problems simultaneously. The degree to which they address each of these problems is

[13] This question only had the binary answer option "Yes" or "No" in the online experiment.

[14] Awareness and implementation problems are diagnosed when respondents select the "Yes" options. An intention problem is diagnosed when respondents select the "No" options.

Figure 4: Precision of the Intervention Effect Prediction - Qualitative Anamnesis



Notes: This figure shows the prediction of intervention effects based on the qualitative anamnesis on the reduction of mistakes compared to the actual effects found for all interventions and settings. We show both the predicted effects and the refined predictions. The displayed p-values refer to differences in predicted effects across settings.

estimated from Table B.7. Hence, the refined prediction for an intervention is the sum of the predicted effectiveness for each problem type, discounted by the degree to which they address the underlying problem. As before, these predictions are then normalized by the baseline's average mistakes.[15]

Results of the Qualitative Anamnesis - Figure 4 shows the qualitative predictions and actual effectiveness across settings. For reminders, we find that the qualitative anamnesis can also effectively be used to predict treatment effects. Predictions for reminders are larger for Setting AWA than in Setting INT and IMP (both p<0.001). Predictions for incentives are larger in Setting INT than in Setting AWA (p=0.002) and Setting IMP (p=0.035).

[15]If a problem is diagnosed, the predicted effects of intervention i are given by $PE_i = \frac{\alpha}{n}$, where α is the actual number of mistakes and $n \in \{1, 2, 3\}$ is the number of diagnosed problems. Refined predictions are calculated by $RefinedPE_i = \sum_{s=1}^{S} PE_i \cdot \delta_s^i$, where δ_s^i is the discount term from Table B.7 indicating how much the intervention resolved the respective problem in setting s. To estimate the relative mistake reduction, we normalize the prediction by the baseline mistakes.

Predictions for simplifications are larger in Setting IMP than in Setting AWA (p=0.001) and Setting INT (p=0.052). Predictions become even more accurate when we use the refined predictions (red crossfades). As a robustness check, we also diagnose a problem type if only one of the two respective anamnesis questions is positive and find that this specification even further improves precision, which we show in Figure A.3 in the Appendix.

We conclude that the qualitative anamnesis also performs well in predicting the actual treatment effects. Appendix Table B.8 confirms the high correlation using the regression statistics.

5.2 Field Application

So far, we have tested our framework (both with the quantitative and the qualitative anamnesis) in a controlled laboratory setting and with an abstract task. As a final part of our paper, we now want to demonstrate the usefulness of our framework also for field settings. In other words, we would like to address the external validity of our approach. For this purpose, we use our anamnesis to diagnose COVID-19 booster hesitancy and match our predictions with actual treatment effects from a study conducted in 2022 by Milkman et al. (2024). Their study tested reminders and free Lyft rides to encourage COVID-19 booster vaccinations in a megastudy with around 3.66 million subjects in the US. Participants were adult CVS Pharmacy patients who resided in one of 65 US metropolitan areas and had received their primary COVID-19 vaccine series, but were not fully boostered. They found that offering people free Lyft rides to pharmacies has no additional benefit beyond sending subjects text messages reminding them to get vaccinated. Experts and laypeople alike failed to predict these treatment effects.

Their study has several features that make it ideal for our purposes: First, it allows us to sample people with very similar characteristics using an online survey. Second, their high-powered study shows a clear positive effect of reminders and a clear, yet unexpected, null effect of offering free Lyft rides. We assume that the reminders were designed to reduce awareness problems, while the Lyft rides were designed to reduce implementation problems.[16] Finally, they found a strong heterogeneous treatment effect. Reminders were three times as effective among those who had received at least one booster in the past.

We ran a pre-registered (https://doi.org/10.17605/OSF.IO/GXQ3C) online study on Prolific in May 2025 to conduct our anamnesis. We matched the sampling criteria as closely as possible to the ones used in the megastudy by Milkman et al. (2024). Participants had to

[16] We follow the reasoning of Milkman et al. (2024), who argue that the free Lyft rides eliminate frictions of transportation hurdles that can hinder the implementation of a vaccination intention.

be adult CVS Pharmacy patients who reside in one of 65 US metropolitan areas selected for study inclusion in Milkman et al. (2024), and had previously received at least their primary COVID-19 vaccination, but had not yet received all available boosters. We conducted our anamnesis with 1,006 new participants.[17]

Qualitative Anamnesis Questions - We elicited problems with receiving a COVID-19 booster by using a 5-point Likert scale on the following questions. As pre-registered, we diagnose individuals to have one of the three problem types if both corresponding anamnesis questions indicate so. To measure awareness problems, we asked: "Did the COVID booster vaccine get overlooked in the hustle of your daily life?" and "Did you simply forget to get one of the COVID booster vaccines in one of the seasons?". To measure intention problems, we asked: "Were you ever determined to get one more COVID booster vaccine in the past?" and "Did you ever plan to get one more COVID booster vaccine in the past?". To measure implementation problems, we used: "I consciously did not get a COVID booster vaccine in the past, although I initially planned to get one," and "I had difficulties getting an additional COVID booster vaccine that I planned to get".

Results - As hypothesized, we indeed find a larger fraction of people diagnosed with only an awareness problem than only an implementation problem (23.7% vs 4.9%, t-test: $p<0.001$). Given that the Lyft ride treatment likely addresses both implementation and awareness problems, we additionally test whether diagnoses of having only awareness problems occur more frequently than diagnoses of implementation problems, whether alone or in combination with awareness problems. The difference is also statistically significant (23.7% vs 11.7%,t-test: $p<0.001$). Hence, based on our anamnesis, we would correctly predict reminders to be much more effective than the provision of free Lyft rides. Figure A.4 shows the full distribution of diagnosed problems.

Next, we test whether we can use our diagnoses to predict treatment effect heterogeneity. Recall that the reminder in the Milkman paper is about three times as effective for boostered individuals compared to those who were not yet boostered. Accordingly, we find that people who were previously boostered have awareness problems that are 2.5 times as high as those without a booster (11% vs. 27.4%, $p<0.001$, Table B.9).[18]

We conclude that our anamnesis and diagnosis can be useful to predict field behavior

[17] We pre-registered 1,000 participants. In order to match the sample characteristics by Milkman et al. (2024), we had to filter out ineligible participants on a rolling basis, which led to an oversampling of 6 participants.

[18] We also pre-registered as secondary hypotheses that, directionally, older subjects and Medicare participants would display more awareness problems. The second Column of Table B.9 shows hardly any effect of age and a negative coefficient of Medicare.

in a setting where experts and laypeople fail to predict the treatment effects. Our results mitigate concerns that our framework can only be applied in very controlled, lab-style settings. Rather, it also performs well in a field setting, despite several empirical features that complicated an accurate diagnosis. First, we collected data 2.5 years later than Milkman et al. (2024), who had run their study in 2022. This means that by now, there are more years in which people could have, e.g., forgotten about one of their boosters in one of the years or failed to get a planned booster, which likely leads to an exaggeration of the diagnosed problems. Second, in 2022, COVID was likely more present in people's minds, meaning that their motivations and attention to get boostered might have been different. Third, most people consider vaccinations in the fall. We collected data in May 2025. As a consequence, participants have to recall the reasons for not getting an additional booster from quite distant time points. Despite these limitations, our diagnosis still reveals that awareness problems outweigh implementation problems, which in turn accurately predicts the greater effectiveness of reminders over free Lyft rides in raising COVID booster uptake.

6 Conclusion

The critical role of diagnosing underlying fundamental problems prior to designing interventions has been recognized in previous literature. Yet, surprisingly few papers offer transparency regarding the rationale behind their choice of intervention within their specific setting. This lack of transparency is notable, especially given the large number of ineffective interventions documented in the literature. We argue that ineffective interventions and the lack of a systematic diagnosis are intertwined. In order to design effective interventions, we need a systematic diagnosis. Hitherto, however, the literature lacks a systematic, generalizable, and parsimonious framework to diagnose fundamental problems and predict intervention effectiveness. Our paper contributes to this gap by introducing an empirically validated framework.

Our approach prioritizes both practicality and parsimony, which ensures broad applicability. Inevitably, this can introduce measurement imperfections due to its reliance on subjects' self-reported recall of past intentions and beliefs. We view our framework as an initial tool to identify the type of fundamental problem. Our framework is not restrictive; where appropriate, it can be easily extended, and potential biases can be rigorously tested using other methodologies. Despite its simplicity, we find that our framework performs well in predicting intervention effects. On average, an ex-ante predicted effectiveness of 10% translates into an actual effectiveness of interventions of 8.92%. Experts, in comparison, tend to largely overestimate treatment effects (Milkman et al., 2021b; DellaVigna & Linos,

2022; Milkman et al., 2024).

While our framework can help to identify the type and extent of the fundamental problem, there are still multiple ways to address awareness, intention, and implementation problems. We believe that cost-effectiveness can be a guiding principle when considering intervention design. For example, when individuals do not intend to work more because of misperceptions regarding their net wage, correcting the misperceptions may be more cost-effective than increasing the actual net wage. For this to be effective, one would have to understand how widespread those misperceptions are. Another related point is that one can imagine that extremely high incentives could solve any problem. For example, offering someone a million dollars to eat an apple a day would likely resolve any related awareness, intention, and implementation problems. However, such an intervention would be far too costly when a simple reminder could also solve the awareness problem to a large extent.

The implications of our findings are far-reaching, particularly in terms of the selection and targeting of interventions. By understanding the underlying fundamental problems, policymakers, researchers, and practitioners can design more effective interventions to address the most pressing challenges of our time.

References

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. In: *Action Control*. Springer Berlin Heidelberg, 11–39.

— (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes* 50(2), 179–211.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics* 130(3), 1117–1165.

Benartzi, S., J. Beshears, K. L. Milkman, C. R. Sunstein, R. H. Thaler, M. Shankar, W. Tucker-Ray, W. J. Congdon & S. Galing (2017). Should Governments Invest More in Nudging? *Psychological Science* 28(8), 1041–1055.

Bhargava, S. & D. Manoli (2015). Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review* 105(11), 3489–3529.

Bordalo, P., N. Gennaioli & A. Shleifer (2022). Salience. *Annual Review of Economics* 14(1), 521–544.

Brody, I., H. Dai, S. Saccardo, K. Milkman, A. L. Duckworth, M. Patel & D. Gromet (2023). Targeting Behavioral Interventions Based on Past Behavior: Evidence from Vaccine Uptake. *Working Paper*.

Bruelisauer, M., L. Goette, Z. Jiang, J. Schmitz & R. Schubert (2020). Appliance Specific Feedback and Social Comparisons: Evidence From a Field Experiment on Electricity Saving. *Energy Policy* 145, 111742.

Bryan, C. J., E. Tipton & D. S. Yeager (2021). Behavioural Science Is Unlikely to Change the World Without a Heterogeneity Revolution. *Nature Human Behaviour* 5(8), 980–989.

Bryan, G., D. Karlan & S. Nelson (2010). Commitment Devices. *Annual Review of Economics* 2(1), 671–698.

Cala, P., T. Havranek, Z. Irsova, M. Luskova, J. Matousek & J. Novak (2025). Financial Incentives and Performance: A Meta-Analysis of Experiments in Economics. *Working Paper*.

Chen, D. L., M. Schonger & C. Wickens (2016). oTree — An Open-Source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance* 9, 88–97.

Datta, S. & S. Mullainathan (2014). Behavioral Design: A New Approach to Development Policy. *Review of Income and Wealth* 60(1), 7–35.

DellaVigna, S. & E. Linos (2022). RCTs to Scale: Comprehensive Evidence From Two Nudge Units. *Econometrica* 90(1), 81–116.

DellaVigna, S. & D. Pope (2018a). Predicting Experimental Results: Who Knows What? *Journal of Political Economy* 126(6), 2410–2456.

DellaVigna, S. & D. Pope (2018b). What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies* 85(2), 1029–1069.

Demeritt, A. & K. Hoff (2018). The Making of Behavioral Development Economics. *History of Political Economy* 50(S1), 303–322.

Duflo, E. & E. Saez (2003). The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *The Quarterly Journal of Economics* 118(3), 815–842.

Engl, E. & S. K. Sgaier (2020). CUBES: A Practical Toolkit to Measure Enablers and Barriers to Behavior for Effective Intervention Design. *Gates Open Research* 3(886).

Falk, A., A. Becker, T. Dohmen, D. Huffman & U. Sunde (2023). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *Management Science* 69(4), 1935–1950.

Fryer, R. G., S. D. Levitt, J. List & S. Sadoff (2022). Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment. *American Economic Journal: Economic Policy* 14(4), 269–299.

Gabaix, X. (2019). Behavioral Inattention. In: *Handbook of Behavioral Economics - Foundations and Applications*. Ed. by B. D. Bernheim, S. DellaVigna & D. Laibson. Vol. 2. North-Holland, 261–343.

Gneezy, U., S. Meier & P. Rey-Biel (2011). When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives* 25(4), 191–210.

Hossain, T. & J. A. List (2012). The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science* 58(12), 2151–2167.

Hoxby, C. M. & S. Turner (2015). What High-Achieving Low-Income Students Know about College. *American Economic Review: Papers & Proceedings* 105(5), 514–17.

Jachimowicz, J. M., S. Duncan, E. U. Weber & E. J. Johnson (2019). When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects. *Behavioural Public Policy* 3(2), 159–186.

Jensen, R. (2010). The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics* 125(2), 515–548.

Kasy, M. & A. Sautmann (2021). Adaptive Treatment Assignment in Experiments for Policy Choice. *Econometrica* 89(1), 113–132.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics* 112(2), 443–477.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Crown/Archetype.

Löfgren, Å. & K. Nordblom (2020). A Theoretical Framework of Decision Making Explaining the Mechanisms of Nudging. *Journal of Economic Behavior and Organization* 174, 1–12.

Maier, M., F. Bartoš, T. D. Stanley, D. R. Shanks, A. J. L. Harris & E.-J. Wagenmakers (2022). No Evidence for Nudging After Adjusting for Publication Bias. *Proceedings of the National Academy of Sciences* 119(31), e2200300119.

Mertens, S., M. Herberz, U. J. J. Hahnel & T. Brosch (2022). The Effectiveness of Nudging: A Meta-Analysis of Choice Architecture Interventions Across Behavioral Domains. *Proceedings of the National Academy of Sciences* 119(1), e2107346118.

Michie, S., M. M. van Stralen & R. West (2011). The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions. *Implementation Science* 6(42).

Milkman, K. L. et al. (2021a). A Megastudy of Text-Based Nudges Encouraging Patients to Get Vaccinated at an Upcoming Doctor's Appointment. *Proceedings of the National Academy of Sciences* 118(20), e2101165118.

Milkman, K. L. et al. (2021b). Megastudies Improve the Impact of Applied Behavioural Science. *Nature* 600(7889), 478–483.

Milkman, K. L. et al. (2024). Megastudy Shows that Reminders Boost Vaccination but Adding Free Rides Does Not. *Nature* 631(8019), 179–188.

Münscher, R., M. Vetter & T. Scheuerle (2015). A Review and Taxonomy of Choice Architecture Techniques. *Journal of Behavioral Decision Making* 29(5), 511–524.

Newell, R. G. & J. Siikamäki (2014). Nudging Energy Efficiency Behavior: The Role of Information Labels. *Journal of the Association of Environmental and Resource Economists* 1(4), 555–598.

Opitz, S., D. Sliwka, T. Vogelsang & T. Zimmermann (2024). The Algorithmic Assignment of Incentive Schemes. *Management Science* 71(2), 1546–1563.

Rockenbach, B., S. Tonke & A. R. Weiss (2025). A Large-Scale Field Experiment to Reduce Nonpayments for Water: From Diagnosis to Treatment. *Review of Economics and Statistics*, 1–14.

Rodrik, D. (2010). Diagnostics before Prescription. *Journal of Economic Perspectives* 24(3), 33–44.

Szaszi, B., A. Higney, A. Charlton, A. Gelman, I. Ziano, B. Aczel, D. G. Goldstein, D. S. Yeager & E. Tipton (2022). No Reason to Expect Large and Consistent Effects of Nudge Interventions. *Proceedings of the National Academy of Sciences* 119(31), e2200732119.

Szaszi, B., A. Palinkas, B. Palfi, A. Szollosi & B. Aczel (2017). A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work. *Journal of Behavioral Decision Making* 31(3), 355–366.

Thaler, R. H. & H. M. Shefrin (1981). An Economic Theory of Self-Control. *Journal of Political Economy* 89(2), 392–406.

Tonke, S. (2025). Shaping Identity: Evidence from a Large-Scale Field Experiment. *Journal of Political Economy Microeconomics*. Forthcoming.

Toussaert, S. (2018). Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment. *Econometrica* 86(3), 859–889.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association* 18(6), 3045–3089.

White, H. (2019). The Twenty-First Century Experimenting Society: The Four Waves of the Evidence Revolution. *Palgrave Communications* 5(47).

Appendix A: Figures

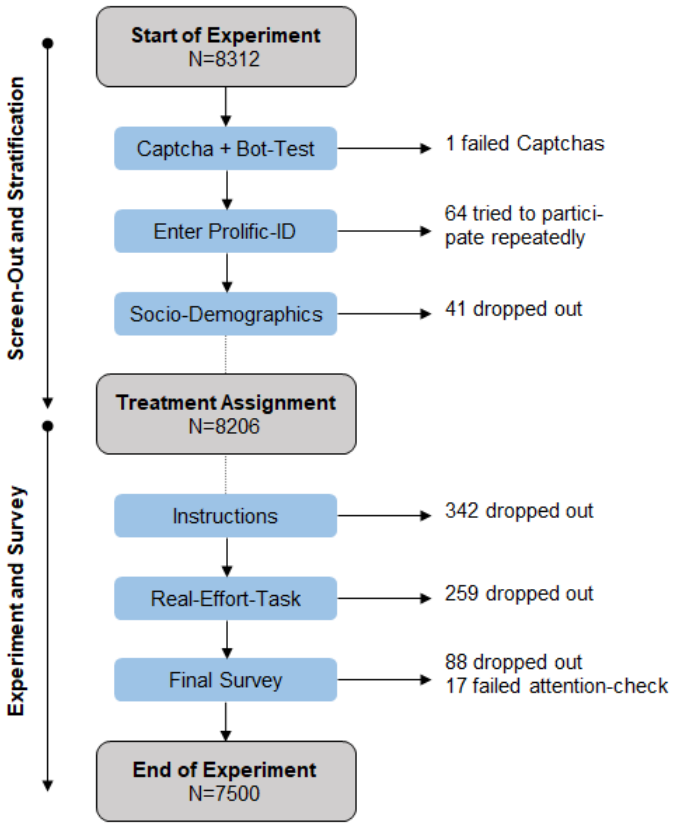


Figure A.1: Flowchart of the sampling process

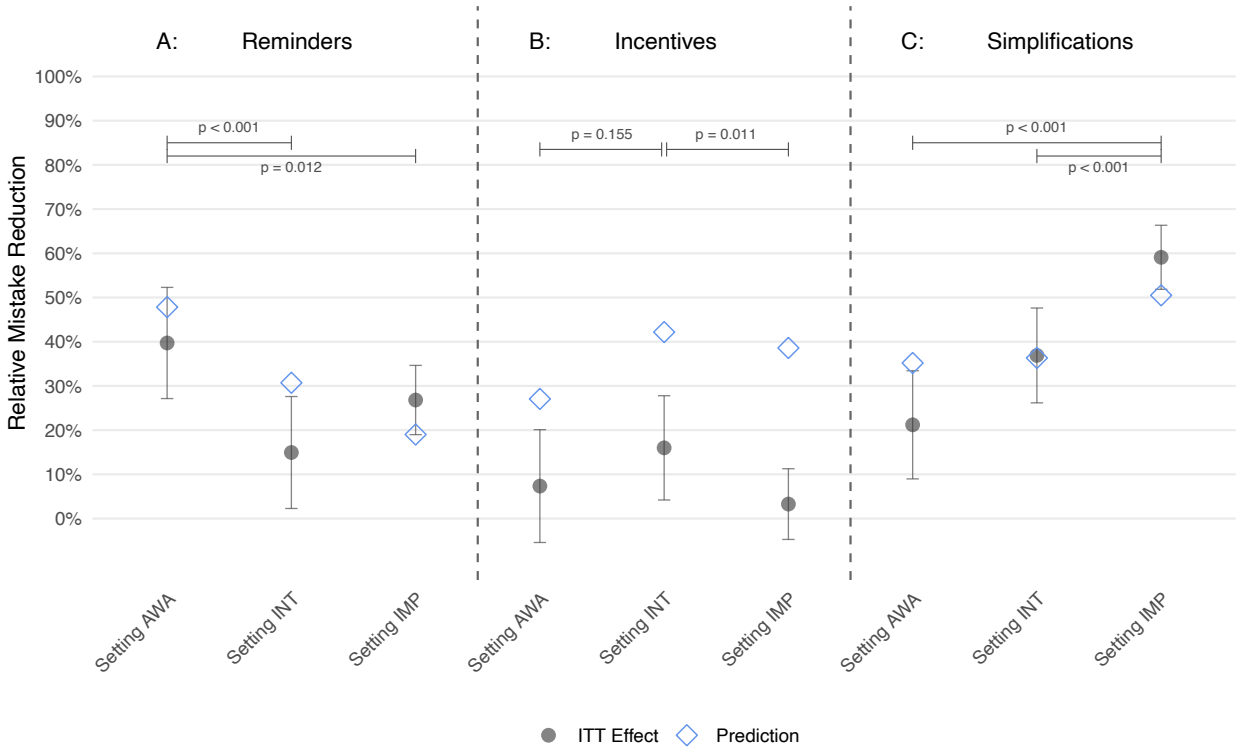


Figure A.2: ITT Effect Heterogeneity Across Settings

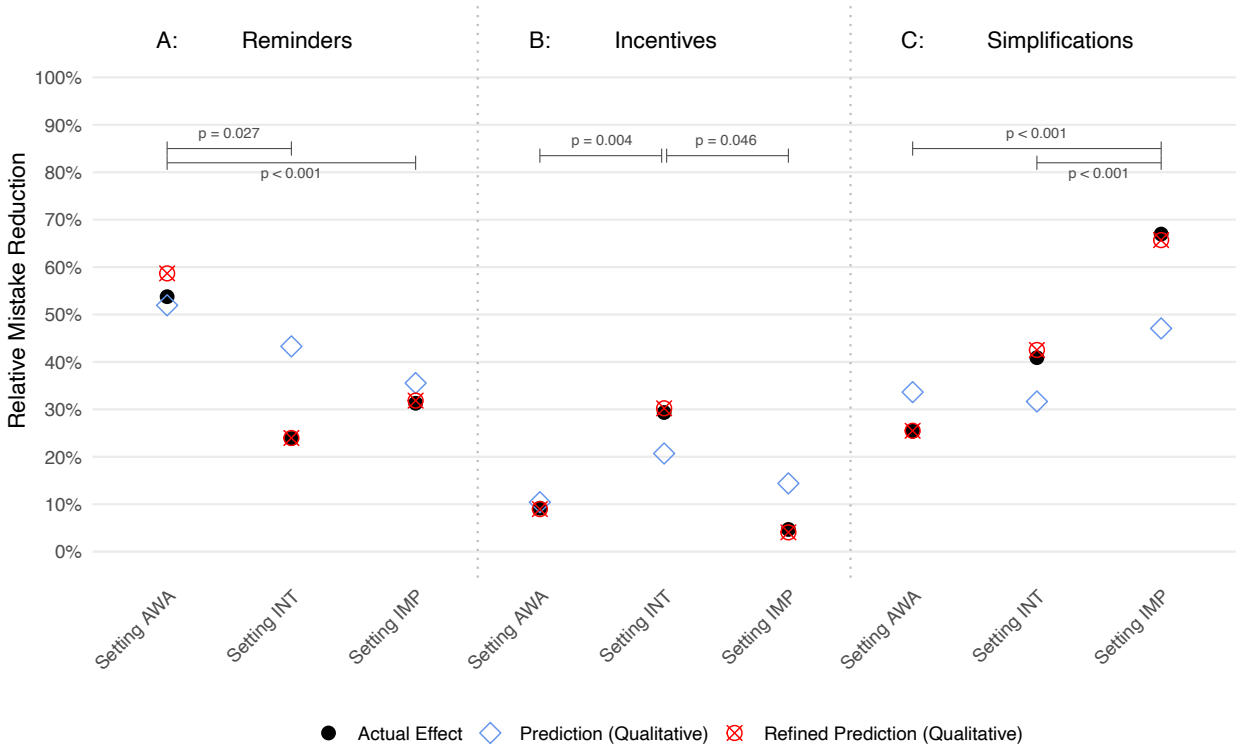


Figure A.3: Precision of the Intervention Effect Prediction - Qualitative Anamnesis (Liberal Classification)

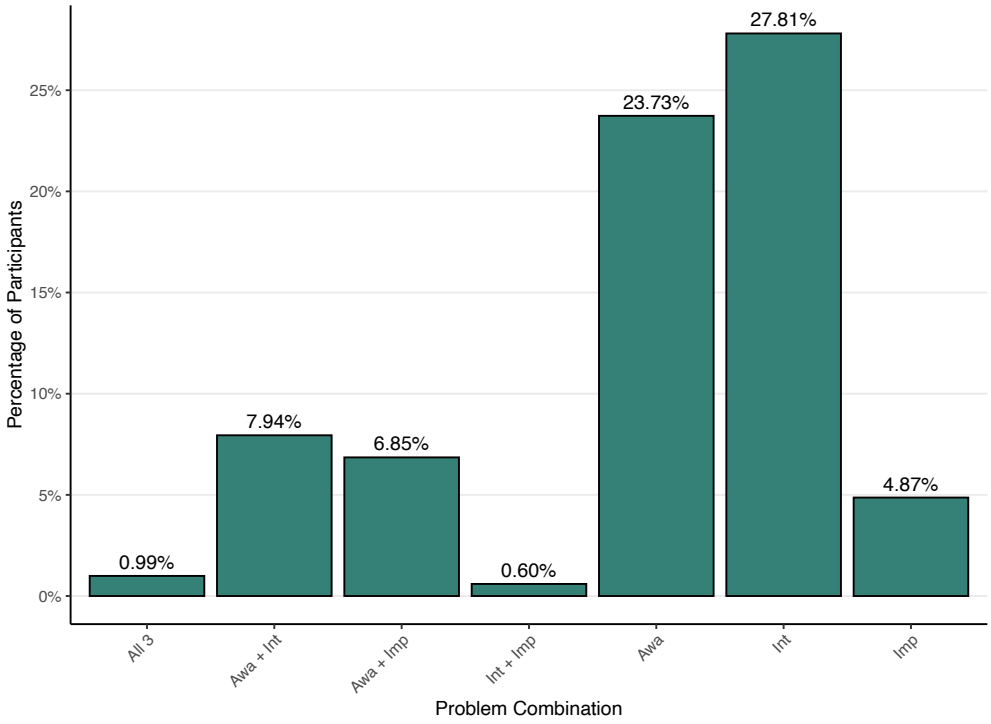


Figure A.4: Diagnosed problems for getting an additional COVID-19 booster

Appendix B: Tables

Table B.1: Balance table for the Experimental sample

Experimental Group	N	Age	Female	College	Republ.	Income	HH-Size	Studying	Unempl.	Working	Self-empl.	Retired
Setting AWA:												
Baseline	636	34.728	0.492	0.627	0.132	8.429	2.789	0.116	0.145	0.615	0.086	0.038
Reminder	629	34.809	0.498	0.622	0.138	8.226	2.801	0.116	0.146	0.595	0.113	0.030
Incentive	634	35.251	0.497	0.618	0.155	8.349	2.793	0.110	0.161	0.585	0.098	0.046
Simplification	640	34.392	0.498	0.630	0.148	8.097	2.694	0.123	0.164	0.597	0.086	0.030
Setting INT:												
Baseline	645	35.101	0.509	0.623	0.146	7.969	2.780	0.096	0.164	0.588	0.112	0.040
Reminder	631	34.778	0.507	0.626	0.166	8.379	2.800	0.105	0.171	0.578	0.116	0.030
Incentive	623	34.920	0.512	0.624	0.141	7.918	2.729	0.109	0.175	0.575	0.109	0.032
Simplification	644	34.991	0.505	0.630	0.148	7.924	2.769	0.115	0.175	0.568	0.102	0.039
Setting IMP:												
Baseline	593	34.642	0.492	0.631	0.123	8.083	2.776	0.115	0.170	0.599	0.081	0.035
Reminder	608	34.243	0.487	0.643	0.148	8.113	2.768	0.130	0.141	0.607	0.084	0.038
Incentive	588	34.350	0.493	0.641	0.107	8.163	2.743	0.121	0.146	0.612	0.095	0.026
Simplification	629	35.134	0.512	0.623	0.138	8.224	2.738	0.118	0.145	0.599	0.113	0.025
F-statistic		0.403	0.178	0.146	1.166	1.740	0.375	0.481	0.805	0.559	1.154	0.739
p-value		0.955	0.999	0.999	0.305	0.059	0.966	0.916	0.635	0.863	0.314	0.702
Significant Diff.		0.000	0.000	0.000	0.106	0.136	0.000	0.000	0.000	0.000	0.015	0.000

Notes: The table shows sociodemographic variables per experimental group. Each row corresponds to one group. The last three rows show the joint F-statistic and the corresponding p-value on whether the means of the groups are significantly different from each other, and the fraction of significant t-tests from all pairwise comparisons. The columns show the averages of the sociodemographic variables per group. *Female, College, Republican, Studying, Unemployed, Working, Self-employed* and *Retired* are indicator-variables that equal 1 if the person meets the corresponding characteristic. *Income* is a categorical variables. The categories 7 and 8 for income correspond to a household net income of \$2,500 to \$3,000 and \$3,000 to \$3,500 per month. The treatment assignment was stratified by age, gender, and college education.

Table B.2: Balance table for the full sample

Experimental Group	N	Finished	Age	Female	College	Republ.	Income	HH-Size	Studying	Unempl.	Working	Self-empl.	Retired
Setting AWA:													
Baseline	682	636	34.930	0.507	0.626	0.130	8.309	2.782	0.116	0.152	0.610	0.085	0.037
Reminder	683	629	35.135	0.504	0.627	0.142	8.221	2.785	0.108	0.149	0.592	0.113	0.038
Incentive	683	634	35.363	0.504	0.625	0.152	8.329	2.792	0.105	0.163	0.592	0.097	0.044
Simplification	683	640	34.524	0.504	0.628	0.146	8.133	2.700	0.123	0.163	0.596	0.085	0.034
Setting INT:													
Baseline	686	645	35.350	0.504	0.625	0.144	7.968	2.770	0.093	0.169	0.589	0.108	0.041
Reminder	679	631	35.047	0.505	0.626	0.163	8.377	2.797	0.103	0.166	0.586	0.112	0.032
Incentive	683	623	35.510	0.505	0.627	0.139	7.971	2.739	0.104	0.176	0.570	0.111	0.040
Simplification	679	644	35.203	0.504	0.627	0.150	7.959	2.766	0.110	0.174	0.571	0.106	0.038
Setting IMP:													
Baseline	685	593	34.978	0.504	0.626	0.121	8.028	2.750	0.111	0.180	0.587	0.085	0.038
Reminder	681	608	34.675	0.505	0.627	0.153	8.060	2.794	0.131	0.151	0.590	0.087	0.041
Incentive	682	588	35.362	0.504	0.629	0.123	7.988	2.748	0.120	0.158	0.591	0.089	0.041
Simplification	683	629	35.120	0.507	0.625	0.142	8.183	2.725	0.113	0.152	0.594	0.114	0.026
F-statistic		7.075	0.345	0.004	0.004	0.865	1.488	0.348	0.692	0.534	0.316	1.189	0.428
p-value		0.000	0.976	1.000	1.000	0.575	0.128	0.975	0.748	0.882	0.983	0.289	0.945
Significant Diff.		0.424	0.000	0.000	0.000	0.030	0.121	0.000	0.015	0.000	0.000	0.000	0.000

Notes: The table shows sociodemographic variables per experimental group. Each row corresponds to one group. The last three rows show the joint F-statistic and the corresponding p-value on whether the means of the groups are significantly different from each other, and the fraction of significant t-tests from all pairwise comparisons. The columns show the averages of the sociodemographic variables per group. *Female, College, Republican, Studying, Unemployed, Working, Self-employed* and *Retired* are indicator-variables that equal 1 if the person meets the corresponding characteristic. *Income* is a categorical variables. The categories 7 and 8 for income correspond to a household net income of \$2,500 to \$3,000 and \$3,000 to \$3,500 per month. The treatment assignment was stratified by age, gender, and college education.

Table B.3: Treatment effects

	Standardized Mistakes					
	Reminders		Incentives		Simplifications	
	(1)	(2)	(3)	(4)	(5)	(6)
Intervention (=1)	−0.538***	−0.539***	−0.293***	−0.292***	−0.670***	−0.666***
	(0.046)	(0.046)	(0.045)	(0.045)	(0.032)	(0.032)
Intervention X Setting INT	0.299***	0.303***			0.261***	0.255***
	(0.061)	(0.061)			(0.035)	(0.035)
Intervention X Setting IMP	0.225***	0.223***	0.247***	0.246***		
	(0.047)	(0.047)	(0.048)	(0.048)		
Intervention X Setting AWA			0.202***	0.204***	0.416***	0.414***
			(0.059)	(0.058)	(0.042)	(0.042)
Age		0.002		−0.001		0.001
		(0.002)		(0.002)		(0.002)
Female (=1)		−0.096***		−0.046		−0.059*
		(0.034)		(0.034)		(0.031)
College Degree (=1)		0.004		−0.059		−0.020
		(0.038)		(0.038)		(0.034)
Republican (=1)		0.055		0.047		0.048
		(0.052)		(0.052)		(0.046)
Net Income		−0.014**		−0.018***		−0.016***
		(0.006)		(0.006)		(0.005)
Household size		0.005		0.026*		0.031**
		(0.013)		(0.013)		(0.012)
City size		0.008		−0.003		0.007
		(0.009)		(0.009)		(0.008)
Working (=1)		0.137***		0.171***		0.077*
		(0.049)		(0.048)		(0.045)
Self-employed (=1)		0.097		0.085		0.004
		(0.071)		(0.070)		(0.065)
Student (=1)		0.049		0.059		−0.005
		(0.067)		(0.068)		(0.061)
Retired (=1)		−0.056		0.089		0.052
		(0.111)		(0.118)		(0.110)
Constant	1.000***	0.927***	1.000***	1.040***	1.000***	0.950***
	(0.026)	(0.102)	(0.026)	(0.097)	(0.026)	(0.093)
Observations	3,742	3,742	3,719	3,719	3,787	3,787
R^2	0.036	0.043	0.010	0.017	0.066	0.072
Adjusted R^2	0.035	0.039	0.009	0.014	0.066	0.068
Residual Std. Error	1.048	1.046	1.042	1.040	0.949	0.947
F Statistic	46.775***	11.823***	12.535***	4.672***	89.459***	20.765***

Notes: This table shows OLS results of the intervention effects on the number of mistakes in the real-effort task across setting with and without control variables based on the participants who finished the experiment. The dependent variable is the number of incorrect answers given in the real-effort task divided by the average number of incorrect answers of the baseline groups per setting for comparison. Which intervention effects to be considered is displayed above the column numbers. *Intervention (=1)* is a dummy variable whose coefficient shows the treatment effect of the respective intervention in the setting that serves as reference group. The interaction terms show the additional effects in the other two settings. Omitted settings for the interaction are the reference groups, respectively. Robust standard errors are in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table B.4: Intention-To-Treat Effects

	Standardized Mistakes					
	Reminders		Incentives		Simplifications	
	(1)	(2)	(3)	(4)	(5)	(6)
Intervention (=1)	−0.397***	−0.395***	−0.160***	−0.162***	−0.591***	−0.587***
	(0.049)	(0.049)	(0.047)	(0.047)	(0.033)	(0.033)
Intervention X Setting INT	0.248***	0.250***			0.222***	0.216***
	(0.064)	(0.064)			(0.040)	(0.040)
Intervention X Setting IMP	0.129**	0.129**	0.127**	0.127**		
	(0.051)	(0.051)	(0.050)	(0.050)		
Intervention X Setting AWA			0.086	0.092	0.379***	0.380***
			(0.061)	(0.061)	(0.047)	(0.047)
Age		0.006***		0.005***		0.004**
		(0.002)		(0.002)		(0.002)
Female (=1)		−0.047		−0.020		−0.036
		(0.033)		(0.033)		(0.031)
College Degree (=1)		−0.011		−0.033		−0.038
		(0.037)		(0.036)		(0.034)
Republican (=1)		0.038		0.011		0.027
		(0.050)		(0.049)		(0.045)
Net Income		−0.019***		−0.025***		−0.016***
		(0.006)		(0.005)		(0.005)
Household size		0.003		0.028**		0.021*
		(0.013)		(0.013)		(0.012)
City size		−0.0004		0.001		0.002
		(0.009)		(0.008)		(0.008)
Working (=1)		0.062		0.066		0.002
		(0.049)		(0.048)		(0.045)
Self-employed (=1)		−0.009		−0.035		−0.066
		(0.069)		(0.069)		(0.064)
Student (=1)		−0.056		−0.037		−0.094
		(0.063)		(0.064)		(0.059)
Retired (=1)		−0.036		0.031		−0.010
		(0.120)		(0.116)		(0.114)
Constant	1.000***	0.946***	1.000***	0.948***	1.000***	0.978***
	(0.024)	(0.098)	(0.024)	(0.093)	(0.024)	(0.092)
Observations	4,096	4,096	4,101	4,101	4,098	4,098
R^2	0.021	0.030	0.003	0.012	0.051	0.057
Adjusted R^2	0.020	0.026	0.002	0.009	0.050	0.054
Residual Std. Error	1.060	1.056	1.041	1.038	0.969	0.967
F-Statistic	28.654***	8.875***	4.251***	3.539***	73.022***	17.680***

Notes: This table shows OLS results of the intervention effects on the number of mistakes in the real-effort task across setting with and without control variables based on the participants who were assigned to an experimental group. The dependent variable is the number of incorrect answers given in the real-effort task divided by the average number of incorrect answers of the baseline groups per setting for comparison. Which intervention effects to be considered is displayed above the column numbers. *Intervention (=1)* is a dummy variable whose coefficient shows the treatment effect of the respective intervention in the setting that serves as reference group. The interaction terms show the additional effects in the other two settings. Omitted settings for the interaction are the reference groups, respectively. Robust standard errors are in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table B.5: Raw Diagnosis Prediction of Intervention Effects

	Actual Mistake Reduction (standardized)			
	Reminders	Incentives	Simplification	Pooled
	(1)	(2)	(3)	(4)
Panel A: Aggregate Level - Intercept				
Raw Predicted Mistake Reduction	0.851***	0.360	2.135***	0.587***
(standardized)	(0.211)	(0.505)	(0.588)	(0.169)
Panel B: Aggregate Level - No Intercept				
Raw Predicted Mistake Reduction	1.085***	0.323***	0.996***	0.759***
(standardized)	(0.082)	(0.060)	(0.051)	(0.038)

Notes: This table shows OLS results of the predicted mistake reduction based on the raw diagnoses of our framework on the actual mistake reduction due to the intervention without adjusting for concurrent problems. We use the diagnosis of the baseline settings in each of the three settings for the prediction of how many mistakes are reduced due to one respective intervention. The actual reduction of mistakes per intervention and setting is used as the outcome variable. Both the predicted and the actual mistake reduction are standardized by the average mistakes of the baseline groups in each setting. Regression coefficients are in percentage terms. For Panel A, we use the the predictions on the individual level as independent variable with robust standard errors in parentheses. To obtain standard errors for the aggregate levels in Panels B and C, we used bootstrapping to resample the original sample 1000 times, calculate the mean intervention and predicted effects for each setting, and perform the OLS analysis on the aggregate data with and without allowing for an intercept. The standard deviations of these bootstrapped coefficients are used as standard errors and reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table B.6: Framework Prediction of ITT Effects

	Actual Mistake Reduction ITT (standardized)			
	Reminders (1)	Incentives (2)	Simplification (3)	Pooled (4)
Panel A: Aggregate Level - Intercept				
Predicted Mistake Reduction	0.521**	0.345	2.091***	0.940***
(standardized)	(0.235)	(0.648)	(0.632)	(0.190)
Panel B: Aggregate Level - No Intercept				
Predicted Mistake Reduction	0.798***	0.250	0.993***	0.704***
(standardized)	(0.234)	(0.291)	(0.188)	(0.095)

Notes: This table shows OLS results of the predicted mistake reduction based on our framework on the actual mistake reduction due to the intervention using the ITT sample and effects. We use the diagnosis of the baseline settings in each of the three settings for the prediction of how many mistakes are reduced due to one respective intervention. The actual reduction of mistakes in the ITT sample per intervention and setting is used as the outcome variable. Both the predicted and the actual mistake reduction are standardized by the average mistakes of the baseline groups in each setting. Regression coefficients are in percentage terms. For Panel A, we use the the predictions on the individual level as independent variable with robust standard errors in parentheses. To obtain standard errors for the aggregate levels in Panels B and C, we used bootstrapping to resample the original sample 1000 times, calculate the mean intervention and predicted effects for each setting, and perform the OLS analysis on the aggregate data with and without allowing for an intercept. The standard deviations of these bootstrapped coefficients are used as standard errors and reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table B.7: Relative Intervention Effects on Diagnosed Problems (Qualitative)

	Change in Diagnosed Problems		
Experimental Group	Awareness (1)	Intention (2)	Implementation (3)
Panel A: Reminders			
Setting AWA	-0.91***	-0.68**	-0.44
Setting INT	-0.77***	-0.02	-0.10
Setting IMP	-0.73***	-0.37**	-0.43**
Panel B: Incentives			
Setting AWA	-0.11	-0.29	-0.21
Setting INT	0.28	-0.81***	-0.50**
Setting IMP	0.09	-0.21	-0.22
Panel C: Simplification			
Setting AWA	-0.09	-0.84***	-0.81**
Setting INT	-0.22	-0.86***	-0.79***
Setting IMP	-0.37**	-0.93***	-0.93***

Notes: The table shows the intervention effects on the extent of diagnosed problems for each experimental setting. All values show the relative change of diagnosed problems relative to the baseline group of the respective setting. Negative values indicate that the extent of the underlying problem was reduced.

Table B.8: Refined Prediction of Intervention Effects by the Framework (Qualitative)

	Actual Mistake Reduction (standardized)			
	Reminders (1)	Incentives (2)	Simplification (3)	Pooled (4)
Panel A: Aggregate Level				
Predicted Mistake Reduction	1.411***	2.141***	2.204	1.111***
(standardized)	(0.338)	(0.895)	(1.453)	(0.211)
Refined Predicted Mistake Reduction	1.044***	2.051***	1.803***	1.262***
(standardized)	(0.270)	(0.704)	(0.383)	(0.104)
Panel B: Aggregate Level - No Intercept				
Predicted Mistake Reduction	0.846***	1.035***	1.219***	1.007***
(standardized)	(0.097)	(0.176)	(0.295)	(0.087)
Refined Predicted Mistake Reduction	1.080***	1.461***	1.438***	1.256***
(standardized)	(0.084)	(0.265)	(0.077)	(0.069)

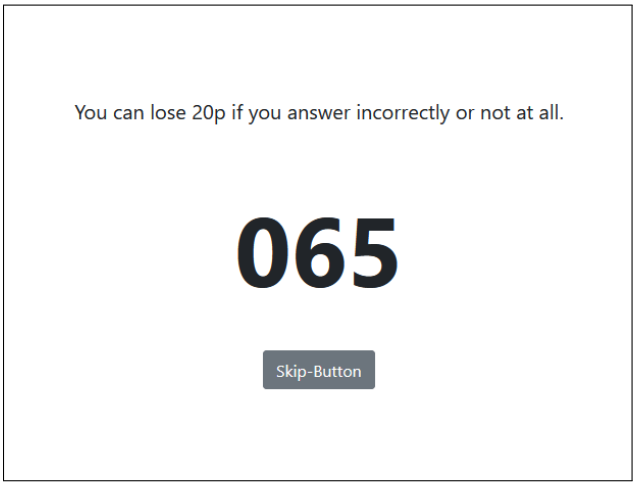
Notes: This table shows OLS results of the refined predicted mistake reduction based on our framework on the actual mistake reduction due to the intervention. We use the diagnosis based on the qualitative anamnesis of the baseline settings in each of the three settings for the prediction of how many mistakes are reduced due to one respective intervention. The actual reduction of mistakes per intervention and setting is used as the outcome variable. Both the predicted and the actual mistake reduction are standardized by the average mistakes of the baseline groups in each setting. Regression coefficients are in percentage terms. To obtain standard errors, we used bootstrapping to resample the original sample 1000 times, calculate the mean intervention and refined predicted effects for each setting, and perform the OLS analysis on the aggregate data with and without allowing for an intercept. The standard deviations of these bootstrapped coefficients are used as standard errors and reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

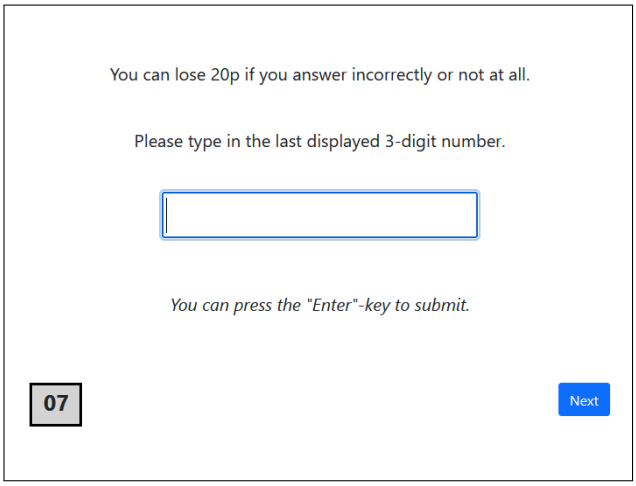
Table B.9: Heterogeneity in the Share of Awareness Problems

	Awarness Problem (=1)	
	(1)	(2)
Boostered (=1)	0.164***	0.167***
	(0.027)	(0.027)
Age (demeaned)		−0.001
		(0.001)
Age^2		0.00003
		(0.0001)
Medicare (=1)		−0.049
		(0.035)
Constant	0.110***	0.111***
	(0.021)	(0.023)
Observations	1,006	1,006
R^2	0.025	0.028
Adjusted R^2	0.024	0.025
Residual Std. Error	0.421	0.421
F Statistic	25.918***	7.311***

Notes: This table shows OLS results of the awareness problem on the pre-registered heterogeneity characteristics. The awareness problem is an indicator equal to 1 if the anamnesis reveals an awareness problem but no other problem with respect to receiving an additional COVID booster vaccine. Boostered equals 1 if the person has already received a booster, and Medicare equals 1 if the person has a medicare health insurance plan. Age is demeaned, whereas the demeaned variable was also used to construct the age-squared term. Robust standard errors are in parentheses.
*p<0.1; **p<0.05; ***p<0.01

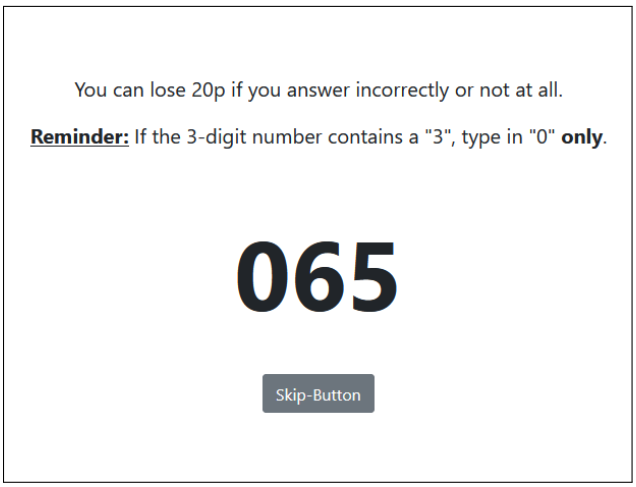
Appendix C: Experimental Screens

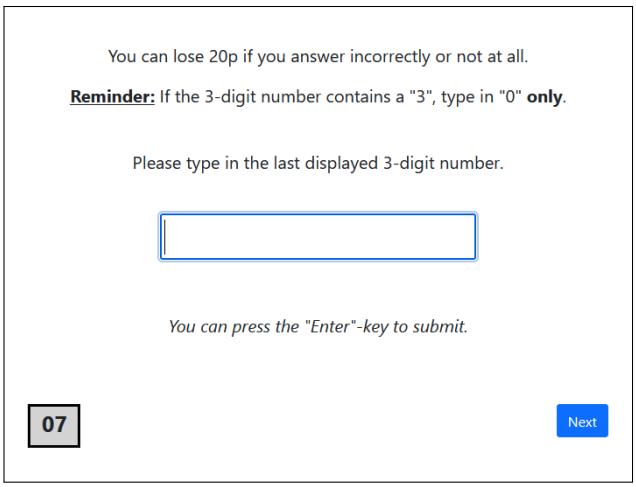




A: Numbers displayed during task B: Query to enter last displayed number

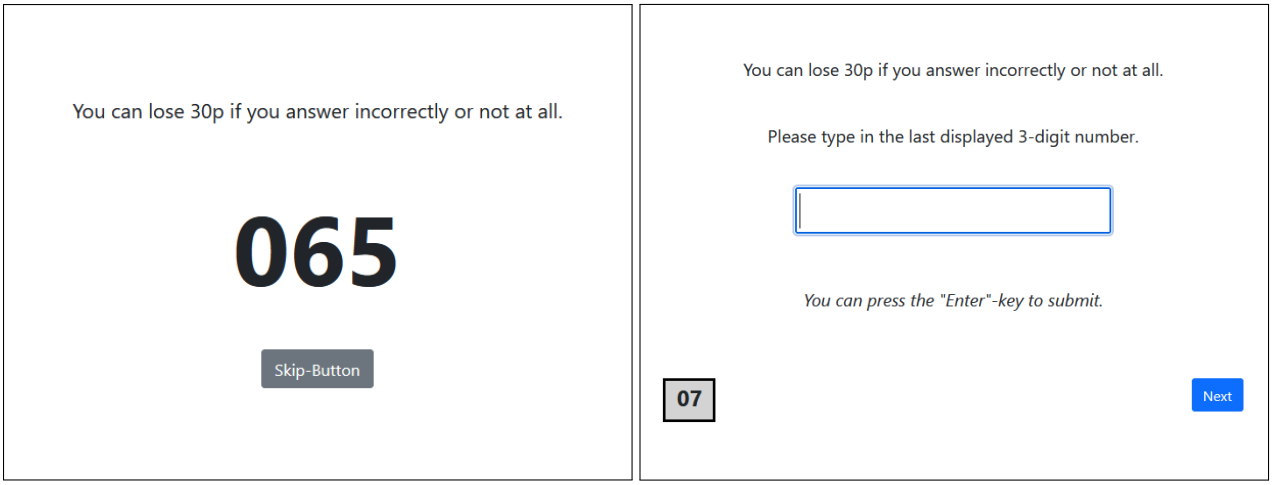
Figure C.1: Screens Setting AWA - Baseline





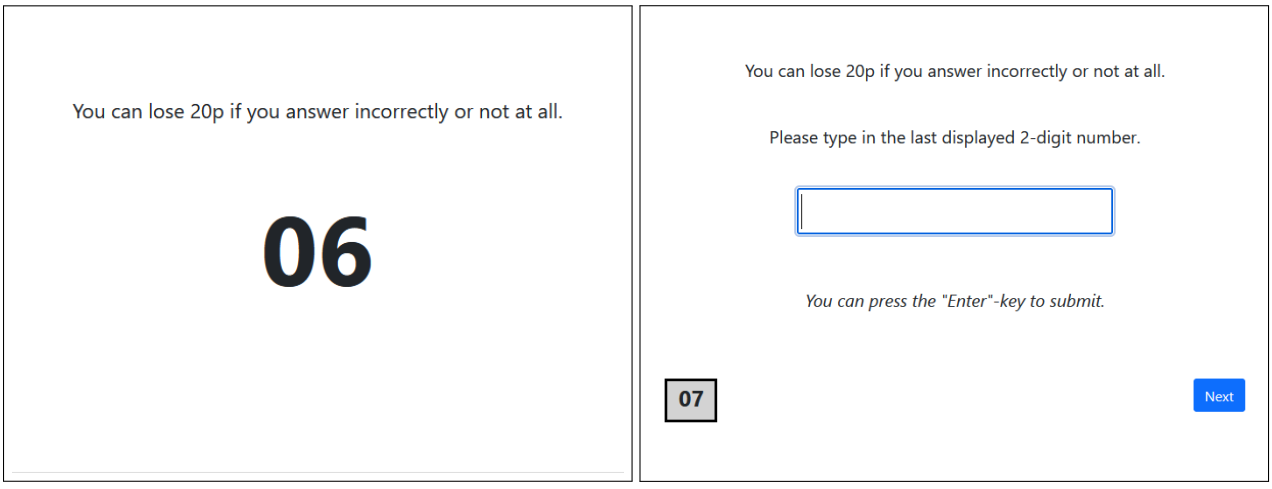
A: Numbers displayed during task B: Query to enter last displayed number

Figure C.2: Screens Setting AWA - Reminder



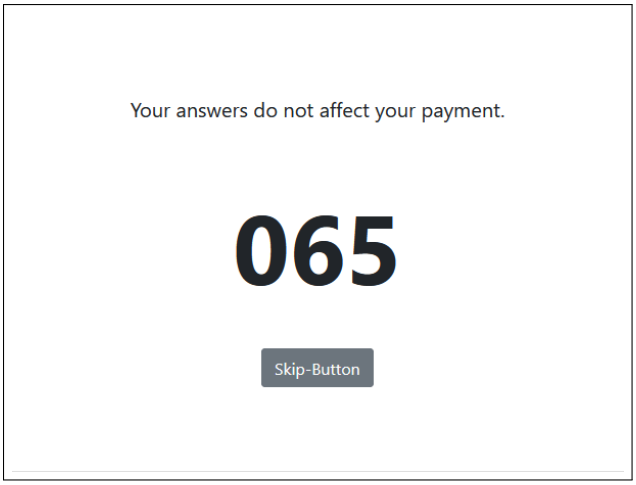
A: Numbers displayed during task B: Query to enter last displayed number

Figure C.3: Screens Setting AWA - Incentives



A: Numbers displayed during task B: Query to enter last displayed number

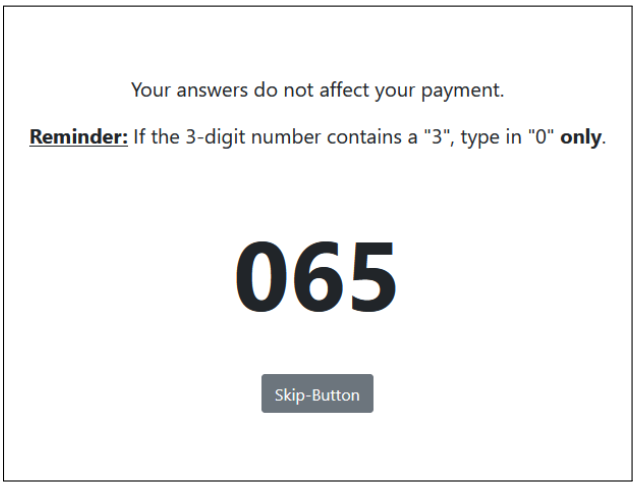
Figure C.4: Screens Setting AWA - Simplifications



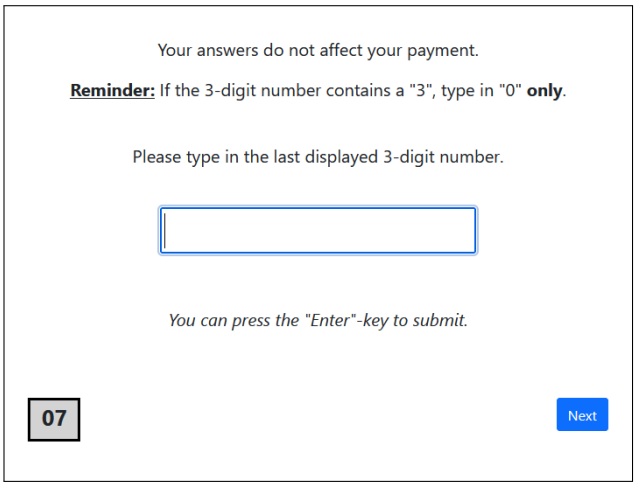
A: Numbers displayed during task

B: Query to enter last displayed number

Figure C.5: Screens Setting INT - Baseline

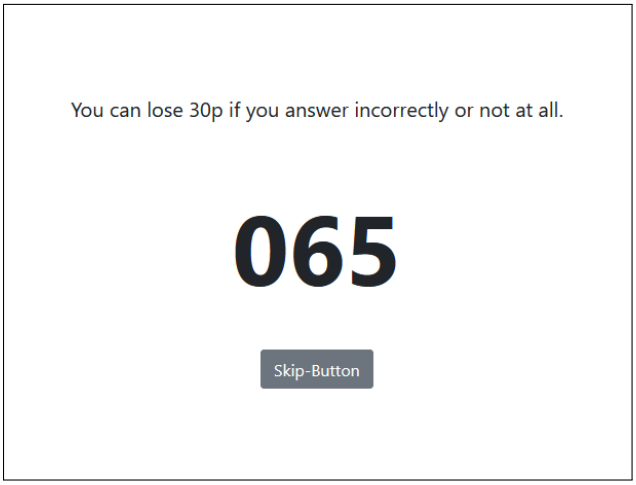


A: Numbers displayed during task

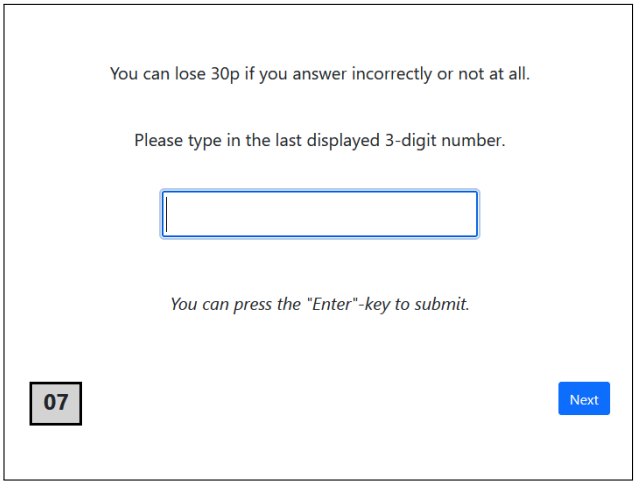


B: Query to enter last displayed number

Figure C.6: Screens Setting INT - Reminder

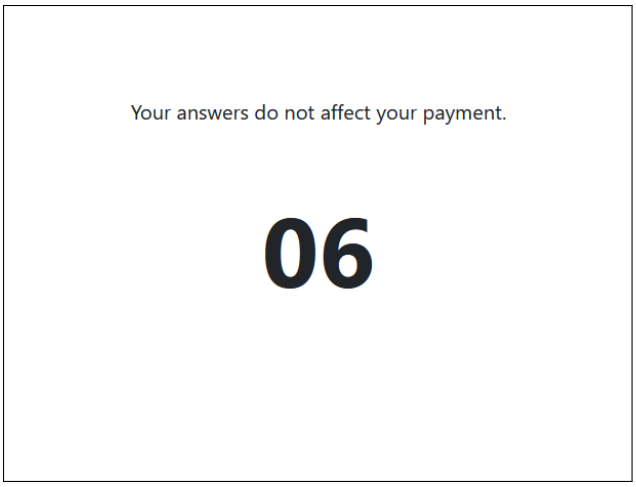


A: Numbers displayed during task



B: Query to enter last displayed number

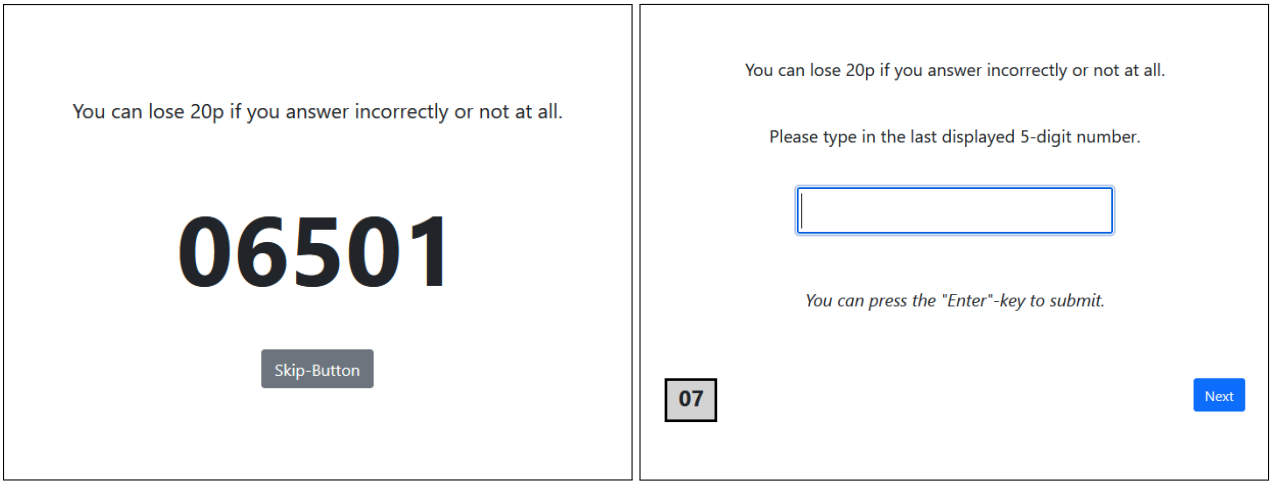
Figure C.7: Screens Setting INT - Incentives



A: Numbers displayed during task

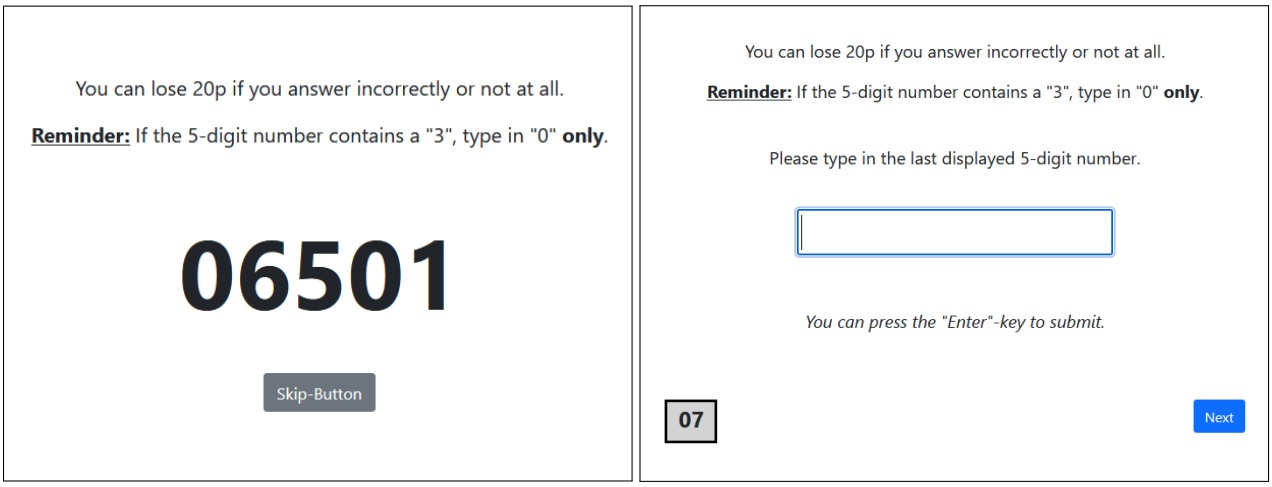
B: Query to enter last displayed number

Figure C.8: Screens Setting INT - Simplifications



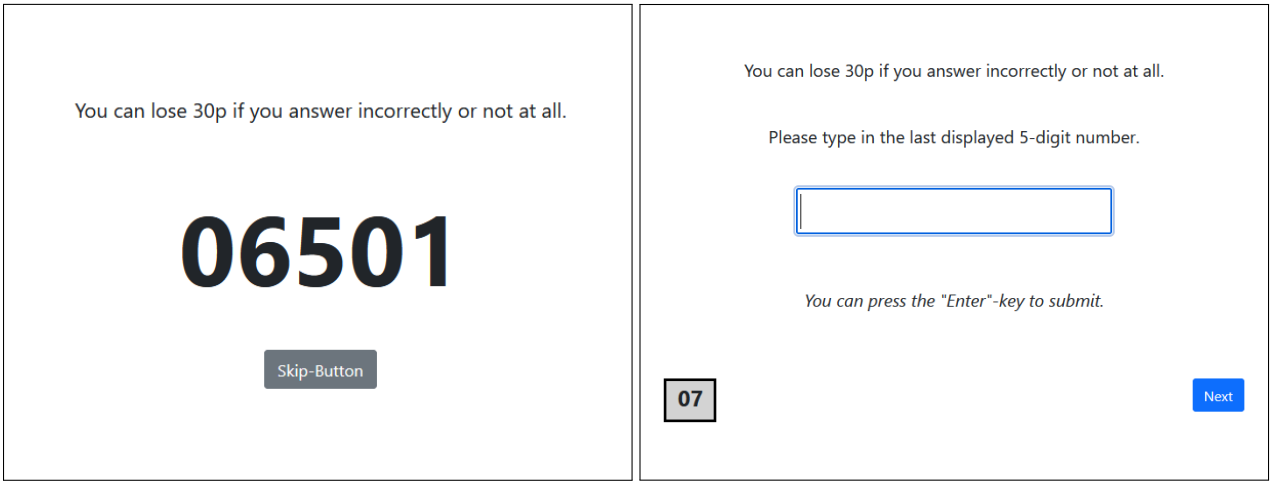
A: Numbers displayed during task B: Query to enter last displayed number

Figure C.9: Screens Setting IMP - Baseline



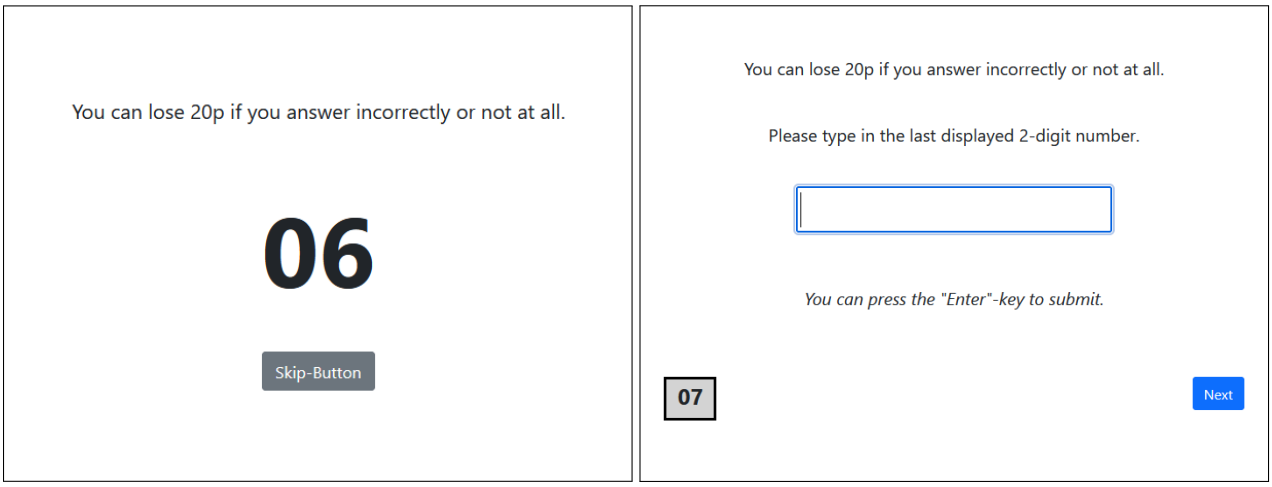
A: Numbers displayed during task B: Query to enter last displayed number

Figure C.10: Screens Setting IMP - Reminder



A: Numbers displayed during task B: Query to enter last displayed number

Figure C.11: Screens Setting IMP - Incentives



A: Numbers displayed during task B: Query to enter last displayed number

Figure C.12: Screens Setting IMP - Simplifications