Empirical Software Engineering (2024) 29:135
https://doi.org/10.1007/s10664-024-10517-w

Guidelines for using financial incentives in software-engineering experimentation

Jacob Krüger[1] · Gül Çalıklı[2] · Dmitri Bershadskyy[3] · Siegmar Otto[4] · Sarah Zabel[4] · Robert Heyer[5,6]

Accepted: 17 June 2024

Abstract

Context: Empirical studies with human participants (e.g., controlled experiments) are established methods in Software Engineering (SE) research to understand developers' activities or the pros and cons of a technique, tool, or practice. Various guidelines and recommendations on designing and conducting different types of empirical studies in SE exist. However, the use of financial incentives (i.e., paying participants to compensate for their effort and improve the validity of a study) is rarely mentioned

Objective: In this article, we analyze and discuss the use of financial incentives for human-oriented SE experimentation to derive corresponding guidelines and recommendations for researchers. Specifically, we propose how to extend the current state-of-the-art and provide a better understanding of when and how to incentivize.

Method: We captured the state-of-the-art in SE by performing a Systematic Literature Review (SLR) involving 105 publications from six conferences and five journals published in 2020 and 2021. Then, we conducted an interdisciplinary analysis based on guidelines from experimental economics and behavioral psychology, two disciplines that research and use financial incentives.

Results: Our results show that financial incentives are sparsely used in SE experimentation, mostly as completion fees. Especially performance-based and task-related financial incentives (i.e., payoff functions) are not used, even though we identified studies for which the validity may benefit from tailored payoff functions. To tackle this issue, we contribute an overview of how experiments in SE may benefit from financial incentivisation, a guideline for deciding on their use, and 11 recommendations on how to design them.

Conclusions: We hope that our contributions get incorporated into standards (e.g., the ACM SIGSOFT Empirical Standards), helping researchers understand whether the use of financial incentives is useful for their experiments and how to define a suitable incentivisation strategy.

Keywords Empirical software engineering · Experimentation · Financial incentives · Study design · Guidelines

Communicated by: Dietmar Pfahl

Extended author information available on the last page of the article

Published online: 10 August 2024

1 Introduction

Empirical studies are important in Software Engineering (SE) research (Juristo and Moreno 2001; Wohlin et al. 2012; Felderer and Travassos 2020; Shull et al. 2008), for instance, to understand the impact of a technique on developers (e.g., using novel testing tools), to investigate relations between properties (e.g., programmers' experience and software defects), or to test theories (e.g., whether agile practices lead to faster releases). For this purpose, various empirical methods can be used, such as controlled experiments, interviews, or questionnaires. Each method comes with its own pros and cons, for example, regarding the trade-offs between internal and external validity (Siegmund et al. 2015; Petersen and Gencel 2013) or between quantitative and qualitative data elicitation (Felderer and Travassos 2020). While particularly challenging to design and conduct, experiments with human participants (Sjøberg et al. 2005; Wohlin et al. 2012; Ko et al. 2015) promise a high degree of internal validity to understand whether, how, and to what degree a property (i.e., independent variable) impacts developers (i.e., in terms of the dependent variable).

One challenge for experiments in SE is the high degree of human factors that are intertwined with software development. Most importantly, inter-individual differences between software developers need to be acknowledged, since selecting just a few participants often leads to a selection bias; meaning that the selected developers represent a specific, not representative subgroup out of all developers (Juristo and Moreno 2001; Höst et al. 2005; Wohlin et al. 2012). The findings of such studies are not generalizable. Consequently, it is crucial to involve a suitable number (i.e., in terms of the population size) of participants who, in addition, must be diverse enough to cover all aspects of inter-individual differences within the overall population. Another challenge are participants who may not complete the tasks in a realistic manner, due to a lack of motivation; even though the experimental design, data collection, and analysis have been well-designed and carefully conducted.

Financial incentives (i.e., monetary compensation) are an established means to address selection bias and motivation issues in various other disciplines, such as experimental economics or behavioral psychology. Such incentives should mimic real-world settings by reflecting developers' situations in practice. For instance, rewarding participants a show-up fee that is derived from developers' wages helps mitigate selection bias, because it can motivate especially higher paid developers to participate. Through *payoff functions* (i.e., mathematical functions relating participants' performance to a payment, cf. Table 1) that address the motivation of developers in the experiment, real-world "motivation scenarios" can be simulated in a more controlled way. Interestingly, various guidelines for empirical SE mention the concept of incentives (Wohlin et al. 2012; Höst et al. 2005; Ralph 2021; Petersen and Wohlin 2009; Carver et al. 2010; Sjøberg et al. 2007). However, to the best of our knowledge, apart from show-up or completion fees that are paid to motivate participation, advanced task-related incentives (i.e., paying incentives based on the actual performance) are rarely employed. Even the ACM SIGSOFT Empirical Standards[1] (Ralph 2021) mention incentives only as (as of September 26, 2022; commit `26815c6`):

1. desirable attribute for longitudinal studies and
2. essential (recruitment) as well as desirable attribute (effect of incentives, improving response rates) for questionnaire surveys.

Note that none of these two mentions in the ACM SIGSOFT Empirical Standards refers explicitly to *financial* incentives. So, it seems that there is a missing awareness of how

[1] https://github.com/acmsigsoft/EmpiricalStandards

advanced financial incentives (i.e., payoff functions) can be used to increase the validity of SE experiments. Unfortunately, missing or misguided incentivisation in experiments may even hamper the validity.

1.1 Goals

With this article, we aim to provide a better understanding and recommendations for using financial incentives in human-oriented SE experimentation, including controlled experiments, quasi experiments, experimental simulations, and field experiments (Wohlin et al. 2012; Stol and Fitzgerald 2020). To this end, we built upon the experiences of other disciplines that have different perspectives on financial incentives (cf. Section 2). First, we consider research from the area of *experimental economics*, an area that relies on laboratory experiments to test theories on human decision-making. In experimental economics, researchers experimentally analyze human decision-making in a variety of cases, including any work-related problems across various domains (e.g., banking, human resources, health). Therefore, experimental economics research that focuses on effort and work or compares different work-related practices and the working environment are relevant for SE, too. Aiming to increase participation in their experiments and improve the validity of the obtained results, researchers in experimental economics often use financial incentives (Harrison and List 2004; Weimann and Brosig-Koch 2019; van Dijk et al. 2001; Erkal et al. 2018). Importantly, incentives in experimental economics are usually more complex than show-up or completion fees that are sometimes used in empirical SE. Besides such fees, researchers in experimental economics often define payoff functions that depend on task correctness (e.g., number of correctly identified bugs), time (e.g., decrease over time spent), or penalties (e.g., for wrongly identified bugs). Second, to reflect on the limitations of financial incentives, especially with respect to the motivation of participants, we also consider research from the area of *behavioral psychology* (Weber and Camerer 2006; Kirk 2013). In this area, many experiments are purposefully designed without task-related financial incentives, since such incentives interfere with intrinsic and extrinsic motivation, and thus can impact (positively as well as negatively) the external validity. By discussing the state-of-the-art on financial incentives in SE experimentation based on insights from these two disciplines, we aim to provide a detailed understanding of the concepts, benefits, and limitations of financial incentives.

1.2 Contributions

We first report the conduct and results of a Systematic Literature Review (SLR) with which we investigated the current state-of-the-art of using incentives in human-oriented SE experimentation. For this purpose, we reviewed 2,284 publications published in 2020 and 2021 at six conferences and five journals with high reputation in SE research. We analyzed 105 publications that report experimental studies with human participants, but only 48 mention some form of incentives, mostly as simple completion rewards. Then, we studied the properties of the individual studies in more detail (e.g., scopes, goals, measurements, participants) to understand whether financial incentives could have been a helpful means to improve their designs. Based on research from experimental economics, behavioral psychology, and our SLR results, we contribute a guideline to decide whether to use financial incentives in SE experiments, derive 11 recommendations to design payoff functions, and exemplify the use of both. Note that we considered the perceptions of two other disciplines (cf. Section 2) to

account for the various different designs of empirical studies—and to understand potential limitations of using incentives that researchers have to keep in mind.

In more detail, we contribute the following in this article:

- We describe how financial incentives are used in two other disciplines to introduce the core concepts, benefits, and limitations (Section 2).
- We report an SLR with which we captured the state-of-the-art of using incentives in SE experimentation (Section 3).
- We discuss the SLR results to understand how financial incentives can help improve the validity of SE experiments (Section 4).
- We define a guideline (Section 5), 11 recommendations (Section 6), and exemplify their use (Section 7) to guide researchers in deciding whether to use and how to design financial incentives in an SE experiment.
- We publish our data in an open-access repository.[2]

Our contributions connect experimental methods used in two other disciplines to SE. Seeing that financial incentives are not well-understood in SE experimentation, and are sparsely used, we argue that we help to mitigate an important gap in existing guidelines for designing experiments in SE. We hope that our contributions help researchers in empirical SE understand trade-offs and design options of financial incentives, and are useful to refine and extend existing guidelines, such as the ACM SIGSOFT Empirical Standards.

1.3 Structure

The remainder of this article is structured as follows. In Section 2, we introduce concepts related to incentives based on established knowledge from experimental economics and behavioral psychology; and discuss the related work. We report the design and conduct of our SLR on incentives in SE experimentation in Section 3. In Section 4, we report and discuss the results of our SLR to provide an understanding of whether and how incentives are used in SE. Then, we build on this discussion to derive a guideline for deciding whether to use financial incentives (Section 5), concrete recommendations for designing financial incentives (Section 6), and exemplify how to use these in SE experimentation (Section 7). Finally, we discuss threats to the validity of our work in Section 8 before concluding in Section 9.

2 Incentives

Incentives can be any form of compensation for the effort participants spend during an empirical study. Typical examples are a set of vouchers that are randomly distributed among all participants of a survey or experiment (Amálio et al. 2020) and non-financial incentives, such as brain scans obtained during fMRI studies (Krueger et al. 2020). In SE research, such incentives are used to increase participation rates for surveys or experiments, and they are usually independent of the actual task performance. While designing an experiment on SE with researchers from experimental economics and behavioral psychology (Krüger et al. 2022), we experienced that particularly in the area of experimental economics financial incentives are used far more systematically. Specifically, researchers in experimental economics design *payoff functions* to increase the validity of experiments. A payoff function is a mapping (i.e., mathematical formula) that defines the relation between participants' choices and their pay-

[2] https://zenodo.org/doi/10.5281/zenodo.12731782

ment (e.g., rewarding correctly solved tasks, penalizing time spent to complete a task), and thus may involve any *task-related* payments (cf. Table 1).

In our experience, SE experimentation is not concerned with, and does not make use of, payoff functions. This personal perception motivated the research we report in this article. Note that we focus on SE experimentation (e.g., controlled experiments, quasi experiments, field experiments), since *payoff functions reward task performance*, which can rarely be done during other empirical SE methods that are used to elicit subjective opinions and experiences (e.g., interviews, questionnaires) or to solve a concrete practical problem (e.g., case studies, action research). Solving a concrete problem is typically connected to the participants' own system, which represents a non-financial incentive (cf. Section 2.4). In the following, we describe financial incentives from the perspectives of experimental economics and behavioral psychology. Particularly, we build on guidelines and research in the area of experimental economics, due to its long history of using financial incentives in laboratory experiments (Harrison and List 2004; Weimann and Brosig-Koch 2019; van Dijk et al. 2001; Erkal et al. 2018).

Table 1 Key concepts of incentives and their definitions. Performance-based (P) payments depend on participants' performance in the experiment, and Task-related (T) payments depend on participants fulfilling their tasks

	term	definition	
		basic concepts	
	Incentive	A driver to motivate participants to perform properly in an experiment, which may be financial (i.e., money or a voucher) or non-financial (e.g., course credits as compensation).	
	Opportunity costs	Benefits lost by choosing one option over the best possible option.	
	Payoff function	A mathematical formula that defines a relation between participants' choices and the respective payments. Payments for individual choices can be positive (e.g., rewarding correctly solved tasks with money) or negative (e.g., penalizing the time needed by subtracting money), and are combined for the final payoff of the function (which is above 0). So, a payoff function can involve all financial incentives that are task-related (see below).	
	Random incentive mechanism	An experimental setup that involves participants facing several tasks with separately defined payoff functions, with only a subset of the tasks being paid out. Participants know about the setup at the beginning of the experiment, but not which tasks are paid out in the end (Cubitt et al. 1998; Baltussen et al. 2012).	
		financial incentives	
$\neg P$	*Show-up fee*	An incentive paid simply for showing up, independently of actual participation. Show-up fees can also be paid to backup participants that do not participate.	$\neg T$
$\neg P$	*Completion fee / lottery*	An incentive paid for completing the experimental tasks, independently of performance (e.g., to all participants, to some participants through a lottery).	T
P	*Winners-take-all tournament*	In such a tournament, the monetary reward is provided to the best-performing participants (e.g., resembling bug bounties) (Cason et al. 2010).	T
P	*Proportional-prize contest*	In such a tournament, the monetary reward is divided among contestants according to their share of total achievement (Cason et al. 2010; Moldovanu and Sela 2001).	T
P	*Piece rate*	Every measurable outcome (e.g., bugs fixed, time taken) is linked to a specified payment or penalty. In contrast to a tournament or contest, the payoff depends only on a participant's own performance (Bull et al. 1987).	T

We provide an overview of the concepts related to incentivizing in Table 1. Finally, we discuss the related work.

2.1 Using Financial Incentives

The most important question when using incentives is: *How do the incentives impact the behavior of participants during an experiment?* Arguably, this is the most complicated aspect of financial incentives, since it is difficult to decide how to incentivize participants to perform "well" during their tasks. Note that "well" in this context refers to whether the incentives are sufficient to induce appropriate preferences, and thus how well the behavior in the laboratory mirrors the behavior in question outside the laboratory.

In his fundamental work on microeconomic systems as an experimental science, Smith (1982) establishes three major conditions for financial incentives in the laboratory that serve as a guide for designing payoff functions to improve experiments' validity and replicability:

> ***Dominance*** means that the incentives are strong enough to overpower other aspects that can motivate the behavior of participants. For instance, consider boredom: If participants stay in the laboratory for some time, they may start feeling bored and could start playing around. The incentives should be strong enough to avoid boredom becoming the main motivator for behavior.
>
> ***Monotonicity*** implies that participants prefer to obtain more of the incentive (e.g., they prefer more money over less).
>
> ***Salience*** defines that a participant's performance in an experiment is transparently linked to the received incentive (e.g., it is clear what reward is paid for correct solutions).

Fundamentally, not much has changed since this initial description of financial incentives (Feltovich 2011). Camerer and Hogarth (1999) discuss whether and when financial incentives matter. While the impact of such incentives on individual experiments can be mixed (i.e., in some cases there are differences, in others there are none), higher incentives usually improve participants' performance, especially for tasks that are responsive to better effort (e.g., mental arithmetic, counting certain letters in a line of text, positioning a slider at the required position). Especially in such cases, incentives should trigger a level of effort that is more similar to the level of effort a person would spend in real-world situations of interest.

In this context, it is important to distinguish different types of financial incentives and tasks, since performance-based incentives (cf. Table 1) can only influence participants' effort within certain limits. By incentivizing outcomes in a performance-based task, participants' effort can only be increased up to their individual maximum abilities (e.g., mental arithmetic). Obviously, participants cannot improve their abilities in a major way during a single experiment. So, Hertwig and Ortmann (2001) conclude in their review that financial incentives improve performance and, even though they do not guarantee optimal decisions, they lead to decisions that are closer to efficient outcomes (as predicted by economic models). Thus, whenever there is no limitation with respect to participants' cognitive ability of solving a task, increasing their motivation leads to a better performance.

Analogous to such findings from economics experiments, insights on survey methods from psychology underpin the role of motivation: The psychological theory of "survey satisficing" (Krosnick 1991) describes participants' strategy in survey situations to answer the questions with the lowest effort possible, which can result in low quality answers. Such a strategy deteriorates the validity of findings, because participants' answers include a component that is connected to the survey only (i.e., answering questions most efficiently in the specific

context of the survey). The theory discusses cognitive processes required for answering surveys and the problem of participants aiming to reduce their own cognitive effort. In a nutshell, the quality of survey responses depends on an interplay of the participants' motivation, their ability, and the difficulty of the task at hand.

To analyze the role financial incentives could have in SE, we focus on one specific set of experiments from experimental economics: experiments involving effort. Such experiments can be designed to implement either *real* or *chosen* effort (Carpenter and Huet-Vaughn 2019). For real effort, a certain task must be exercised, which is linked to the payoff function (cf. Table 1). The pro of such a strategy is a higher external validity, or at least more mundane realism. The con of such a strategy is that some important variables (e.g., costs of effort, intrinsic motivation) are not observed. This issue can be mediated through diligent randomization of participants between treatments. Another issue with real effort is that it is difficult to calibrate the appropriate payoff function. In contrast, chosen effort means the participant chooses a level of effort that directly corresponds to certain monetary costs. Simplified, they face a specific scenario for which they have to decide how much time they would be willing to spend. For example, the chance of identifying a bug per minute is 50% and awards $1, but the costs of searching incrementally increase from $0.1 to $1. Such a design offers a higher level of control at the cost of realism (Carpenter and Huet-Vaughn 2019). Note that we focus on real-effort experimental designs, since only these are concerned with participants exercising a measurable task.

2.2 Benefits of Financial Incentives

Using feasible financial incentives in experiments promises several benefits (B), for instance:

(B_1) improving the participation rate of experiments;
(B_2) improving the realism of an experiment;
(B_3) improving the motivation of participants to exercise the experimental tasks appropriately;
(B_4) reducing the variance in outcomes due to the dominance condition limiting the impact of other motivators—thus, also improving replicability (Camerer and Hogarth 1999); and, via these four,
(B_5) improving the validity (i.e., internal when tasks involve effort, external when representing the real-world).

The sizes of payoffs are typically oriented towards the real-world *opportunity costs* (cf. Table 1) of the participants (i.e., exhibiting similar properties in terms of fixed and variable payoffs or penalties). For example, a payoff to compensate for the time that participants have to spend to finish a task can be defined by considering the average monthly (or hourly) wages of participants in the country in which an experiment is conducted (Harrison and List 2004). Next, we discuss the benefits of improving participation B_1 and realism in experiments B_2. We do not discuss the other benefits (again), since we described how the major conditions of financial incentives, particularly dominance B_4, can improve the motivation of participants B_3 in Section 2.1, and a higher validity B_5 is the consequence of achieving the other four benefits.

Improving Participation The most common argument for using financial incentives is to improve participation, which can help tackle two issues. First, empirical studies require a certain minimum number of participants to ensure the validity of the obtained results (e.g., ensuring statistical power). A large body of evidence from psychology supports the assumption that financial incentives increase response rates in surveys; specifically, they

seem to more than double the odds for responses (Edwards et al. 2005; David and Ware 2014), which is transferable to experiments. Second, selection biases should be mitigated. For this purpose, the laboratory conducting the experiment must have an appropriate reputation and well-defined policies to attract participants. In particular, participants who register for experiments must know a priori that they will be reimbursed for the time they spend during an experiment, independently of what the precise scope of the experiment is. The actual payment can depend on a combination of show-up fees, task-related rewards, penalties, or chance (e.g., winners-take-all tournament, lottery) to mimic different real-world scenarios (Weimann and Brosig-Koch 2019). One example in SE could be to reward or penalize the identification of correct or wrong feature locations, respectively; or to fix faulty configurations. Note that participants should not end up with a negative payment for the whole experiment.

Being aware that their participation in an experiment will be reimbursed makes the setup for participants comparable to methods for incentivizing surveys. For instance, participants are incentivized to fill out surveys to avoid selection bias. Selection bias may occur if participants mainly consist of those who are interested in the survey topic, have a positive attitude towards surveys, or score high on traits, such as openness and pro-socialness (Brüggen et al. 2011; Keusch 2015; Marcus and Schütz 2005). Also, incentives can increase response rates for participants with lower socioeconomic status or of younger age (Simmons and Wilmot 2004). While payoffs in psychology are often based on a lottery, it is established in experimental economics that all participants receive payoffs. Researchers in experimental economics further enhance the use of incentives by questioning how different types of payments (e.g., risky and task-related payments versus fixed show-up fees) can cause selection bias with respect to the risk preferences of participants (Harrison et al. 2009).

Improving Realism Another set of arguments for using financial incentives is based on a general issue of laboratory experiments. Usually, such experiments take place in an environment that differs from the targeted environment in several ways (e.g., by isolating individuals in laboratory cubicles, by using specific measurement equipment like eye-trackers, or by requiring/forbidding the use of specific tools). Consequently, while laboratory experiments excel at improving internal validity, they are usually limited in terms of external validity. Still, it is recommended to maximize the external validity of an experiment, as long as the internal validity is not harmed. Since the majority of experiments in experimental economics

1. focuses on testing economic theories (e.g., game theory),
2. builds on maximization assumptions (i.e., participants aim to maximize their payoff), and
3. concerns problems involving money and time (or tradeable goods that can be easily converted to money),

it is reasonable to conduct experiments in a setting that is comparable to the real world (Hertwig and Ortmann 2001). So, introducing financial incentives to mirror the outside-the-lab situation (e.g., paying professionals) can increase the external validity of an experiment (Schram 2005). Similarly, SE research is heavily concerned with practical problems, and simulating the real world in experiments is an important concern. For instance, researchers may test the theory whether their new technique helps developers detect more bugs (1.) in a shorter period of time (3.), for which they can use financial incentives to motivate real-world behavior (2.).

2.3 Limitations of Financial Incentives

Even though it can be beneficial to use financial incentives in an experiment, researchers must balance these benefits against several limitations. In the following, we discuss two

more pragmatic (e.g., handling costs, controlling participation) and two more fundamental (i.e., addressing habituation, scoping the payoff) limitations. We aim to help researchers understand and resolve such limitations.

Handling Costs Obviously, (financial) incentives increase the costs of an experiment, especially if they are oriented towards real hourly wages. Therefore, when designing an experiment, researchers should consider to what extent the outcome depends on the participants effort and motivation. Considering incentives in relation to the overall costs of a research project, incentives could be considered marginal. For example, a rather large laboratory experiment with about 300 participants and an average payoff per person of around \$20 will cause costs for incentives of roughly \$6, 000. Nonetheless, funding incentives for an experiment can become a challenging issue, depending on the availability of project funds. In this regard, our guidelines can help researchers to plan and reason for such funding within their grant proposals.

Controlling Participation Incentives are helpful for targeting certain groups of participants and reducing sampling bias. However, besides these desired effects, incentives in online surveys can also attract participants who simply click through the survey in order to receive the payment, or even bots. Similarly, online as well as in-person experiments face the problem that participants may only be interested in receiving the payment, without concern for the actual task. Consequently, stronger control is required, especially when incentives are provided as show-up fees. Also, there are different methods (e.g., CAPTCHAs, different types of questions, plausibility checks) that can decrease the threat of bots in online settings (Aguinis et al. 2021). Lastly, a payoff function can also minimize these problems by granting lower payment for random responses—and our guidelines are intended to help researchers define such functions.

Addressing Habituation Singer et al. (1998) found that incentives can raise participants' expectations regarding survey incentivisation in general, but it is not clear how these expectations impact participants' behavior and the quality of their responses in future studies. At worst, habituation processes (i.e., decreases in response strength due to practice (Thompson and Spencer 1966)) could occur. For incentives, habituation means that, over time, their positive effects on the outcomes would dissipate, since participants become accustomed to receiving incentives. In the long run, this effect could increase costs without improving quality to the same extent. Fortunately, the few studies that have been conducted on habituation do not confirm such an effect (Pforr 2015). Esteves-Sorenson and Broce (2020) speculate that unmet payment expectations, which can occur if incentives are provided in one study and withdrawn in a second one, could harm output quality. Laboratories in experimental economics address unmet payment expectations and habituation by having internal quality guidelines including the size of expected payments, and by using payoff functions that put more weight on (dominating) task-related payments than show-up fees. Again, our guidelines can help researchers define such setups for their laboratories and experiments.

Scoping the Payoff Finally, there is a lack of evidence regarding what constitutes the "right" amount of incentives. For instance, findings from experimental economics indicate that the pure presence of financial incentives is rarely the motivating factor for participants, but it is rather their magnitude (Gneezy and Rustichini 2000; Parco et al. 2002; Rydval and Ortmann 2004). This issue broadly refers to the condition of dominance and the so-called *crowding out* of motivation. The term crowding out refers to the observation that extrinsic motivation (e.g., financial incentives) may replace intrinsic motivation, making the total effect on performance more ambiguous (Deci 1971; Frey 1997). For example, Murayama et al. (2010) illustrate on a neurological level how financial incentives can undermine intrinsic motivation when the task has intrinsic value of achieving success. This discussion also led to neurological evidence

indicating that real choices activate reward regions in the brain more strongly and broadly than hypothetical choices (Kang et al. 2011; Camerer and Mobbs 2017).

In addition, ethical concerns have been raised that too large participation incentives may force participants to participate in a study, which is contrary to the principle of voluntariness (Pforr 2015). However, in most social surveys, incentives offered are not that high, and thus unlikely to inappropriately influence participants in terms of, for instance, accepting a higher risk of personal data being disclosed (Singer and Couper 2008). For surveys in behavioral psychology, studies found that financial incentives increase the response rate (and more than non-financial incentives), but they do not seem to improve the quality of responses when considering item-nonresponse (i.e., missing information/values regarding variables) as indicator (Singer and Ye 2013). So, the following question remains unanswered: What should the amount of incentives be to have a positive impact on the response quality? We intend our guidelines to help SE researchers tackle this question for their experiments.

2.4 Related Work

In the following, we provide an overview of the related work, particularly with respect to research on financial incentives in computer science and SE.

Financial versus Non-Financial Incentives The study by DellaVigna and Pope (2018) is probably the most rigorously conducted one on the impact of (non-)financial incentives on participants' motivations. Among others, the authors conducted a large real-effort experiment with 18 treatment arms and over 9,800 participants to analyze different motivators (i.e., financial, behavioral, and psychological). DellaVigna and Pope (2018) illustrate that financial incentives work better than psychological ones. Note that the findings require caution, since they refer only to the specific context of the experiment. Still, Esteves-Sorenson and Broce (2020) obtained similar results when they reviewed over 100 studies on crowding out and conducted their own field experiment. They found that financial incentives did not lead to a crowding out of motivation for intrinsically motivated individuals. However, they state that unmet payment expectations may influence the output quality. Given the variety of tasks as well as structural differences among them (e.g., with respect to the amount of intrinsic motivation), it is necessary to provide tailored incentivisation for a specific task—especially for experiments including real effort. In this context, effort (i.e., the decision on how much effort to put in a task) refers to the way participants can earn money in an experiment.

Financial Incentives in Computer Science We have been aware of a few publications that are concerned with financial incentives in empirical SE. However, to improve our confidence that we did not miss any important studies or guidelines on financial incentives, we performed an automated search on dblp[3] and Scopus.[4] For dblp (last updated July 30, 2021), we used the search string `financial incentive`, which helped us identify publications in computer science that refer to those two terms in their bibliographic data (e.g., title). For Scopus (last updated September 26, 2022), we used the search string `"financial incentive" AND experiment` on titles, abstracts, and keywords; and excluded any subject areas that are not computer science. We obtained 50 and 26 publications, respectively, with some overlap between the two sets. These publications are concerned with topics like using (financial) incentives to motivate online reviews (Wang and Sanders 2019; Burtch et al. 2018), knowledge sharing in social networks (Kettles et al. 2017), or crowdsourcing (Ho et al. 2015; Shaw

[3] https://dblp.org/

[4] https://www.scopus.com

et al. 2011). Such topics are not connected to our goal (i.e., using financial incentives in experiments), or have financial incentives as an inherent property (i.e., crowdsourcing). Consequently, they are not within the scope of our study and intended guidelines, since we are interested in experiments that use financial incentives as a methodological means to simulate realism and reward cognitive effort. Still, the following seven publications are closely related to our own work.

Sjøberg et al. (2005) surveyed controlled experiments in software engineering from 1993 to 2002. While they focused on various other properties of the experiments, Sjøberg et al. (2005) also collected data on the recruitment of participants and the rewards used. Overall, the use of incentives was explicitly mentioned even fewer times compared to our SLR (23 of 113 versus 50 of 105 experiments; cf. Table 2). Similarly, only three experiments in the survey by Sjøberg et al. (2005) reported on incentivizing with money, whereas we found 30 of such experiments (i.e., money or monetary vouchers). Comparing both studies may seem to indicate that there has been a continuous raise in the use of financial incentives, but the covered venues are different and the number of experiments in SE has increased. Consequently, the picture may be skewed. Moreover, the general insights remain identical between the two reviews: financial incentives are somewhat used, but mostly to motivate participation (e.g., using completion fees) whereas payoff functions seem unused. Besides such similarities, our SLR differs considerably from the survey by Sjøberg et al. (2005) due to the focus on financial incentives and coverage of a more recent period. For this reason, the guidelines and recommendations we derive are also completely new.

Glasgow and Murphy (1992) report an experiment with a small software development team in which financial incentives have been used. They found that financial incentives can have negative impact in practice, for instance, reducing social interactions between developers or causing a feeling of injustice. This study highlights that it is challenging to implement the right financial incentives, in practice and in experiments. However, this report is rather old and the details are vague. Instead of practice, we are focusing on motivating participants during experiments, we discuss how to balance the pros and cons of financial incentives, and build on more advanced research on incentives.

Rao et al. (2020) compare different incentivisation schemes aiming to receive deep bug fixes rather than shallow ones. Their study focuses on ways of distributing financial incentives in software markets (e.g., based on what strategies, when, and to whom bug bounties are paid). However, the study is not concerned with experiments, and provides only a simulation of the defined payoff functions. So, this work is complementary to our research, in which we provide actionable contributions for SE researchers for designing experiments with human participants.

Fiore et al. (2014) report on four experiments in which they compared how regular and "surprise" financial incentives impact the participation rate in online experiments. They found that surprising participants with financial incentives yields lower participation rates compared to motivating the incentives from the beginning (i.e., in the advertisement)—but increasing the financial incentive surprisingly after following the advertisement yields even higher participation rates. The results underpin that financial incentives can help increase participation rates in SE experiments. However, we are focusing on how to improve the motivation during the experiment to improve its validity.

Grossklags (2007) discusses the use of financial incentives in experimental economics and exemplifies related studies in computer science. While we also build upon knowledge from experimental economics, there are key differences to our work: In contrast to Grossklags, we (i) systematically elicit the current state-of-the-art of using incentives in SE experimentation

in Section 3; (ii) do not solely focus on experiments on economics, for which incentives are of primary concern (e.g., Grossklags (2007) exemplifies trading in electronic markets or game theory for computer networks); and (iii) derive guidelines for using financial incentives in SE experiments with human participants.

Höst et al. (2005) study non-financial incentives that are the result of the properties of the experimental object. Namely, the authors' findings indicate that if the experimental object is an isolated artifact (e.g., a random piece of code), the validity of the experimental results relies on the participants' will and pride to perform their task correctly. Other factors that incentivize participants may be a code of conduct that open-source developers adhere to or the possibility to improve their grading for student participants. Similarly, presenting the experimental object within an real-world setting can improve participants' motivation. Most prominently, participants' motivation will arguably be the highest if the experiment or field study is conducted within a software project they are working on. For example, conducting an experiment, field study, or action research on code inspection on a real system with the corresponding developers would incentivize participants, since they should be interested to remove the bugs anyway. Consequently, real-world settings are more incentivizing for participants than laboratory examples, which may feel more like a waste of time for the participants. In a related manner, we are concerned with using financial incentives to mimic real-world settings to improve participants' motivation.

Mason and Watts (2009) discuss the relationship between incentivizing experiments conducted on Amazon Mechanical Turk and the participants' performance. For this purpose, they discuss the results of two experiments and confirm findings from experimental economics and behavioral psychology. Namely, their results suggest that financial incentives improve the quantity of tasks performed, but not their quality; and that different forms of incentives (e.g., piece rate versus quota scheme) could significantly impact the quality. However, this work is not concerned with typical SE experimentation (i.e., the tasks were to order images and to solve a word puzzle), but how incentives may impact participants on crowd-sourcing platforms (i.e., Amazon Mechanical Turk). In contrast, we aim to provide guidelines on how to financially incentivize actual SE experiments.

3 A Review of Incentives in SE

Within this section, we report our SLR on experiments and observational studies in SE, for which we followed the guidelines of Kitchenham et al. (2015). In Fig. 1, we provide an overview of our overall process. Please note that we did not review the state-of-the-art in experimental economics because financial incentives (1) are a de-facto standard in this domain; (2) included in established guidelines (Weimann and Brosig-Koch 2019); and (3) even required by many journals, like *Experimental Economics*.[5] Consequently, there is no need to review the state-of-the-art, since we could build on well-established references and extensive research from this domain—in contrast to SE.

3.1 Goal and Research Questions

As motivated, the main purpose of our SLR was to understand the state-of-the-art of how (financial) incentives are used in SE experiments. Besides eliciting evidence in favor of

[5] https://www.springer.com/journal/10683/aims-and-scope:"However, we only consider studies that do not employ deception of participants and in which participants are incentivized."

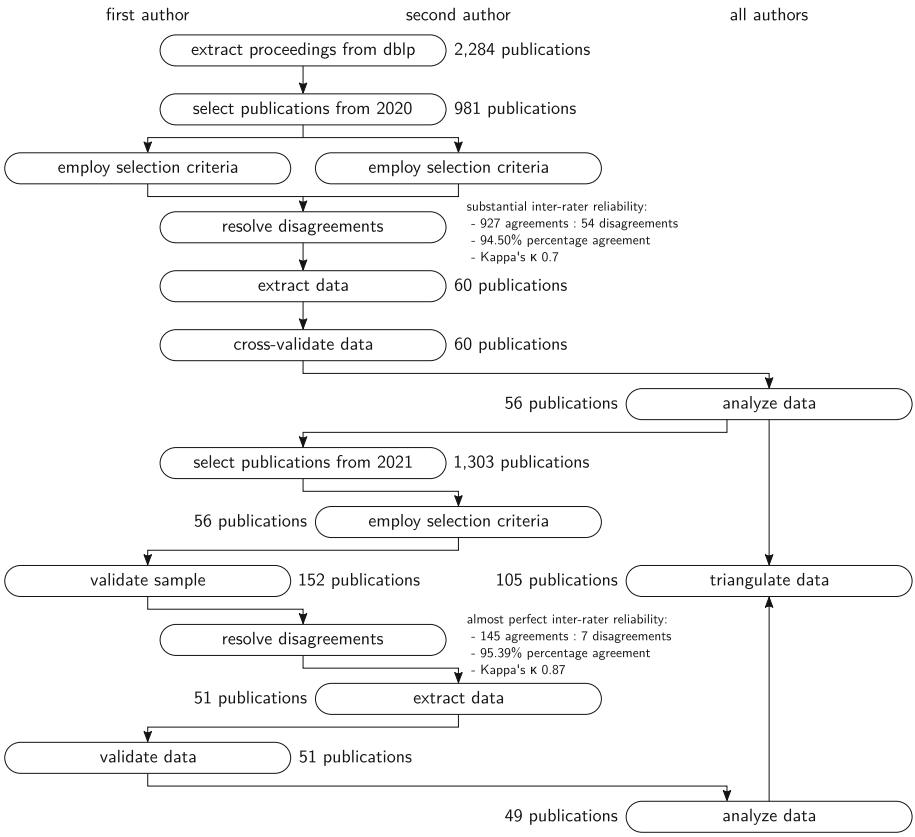

Fig. 1 Overview of the methodology we employed

or in contrast to our subjective perception of publications and existing guidelines on SE experimentation not involving advanced financial incentives, the results should provide us with the information we needed for our interdisciplinary analysis. Particularly, the results should provide the foundation for deriving recommendations for guidelines that are feasible for SE (e.g., considering open-source developers). For this purpose, we had to understand how SE experiments are currently designed, set up, and reported. Using this information, we could identify gaps between SE and other disciplines, allowing us to construct our guidelines and recommendations for using financial incentives.

We defined two research questions to understand to what extent and in what forms financial incentives are used in SE:

> RQ_1 *To what extent are financial incentives used in experiments?*
> We collected the publications to understand whether and to what extent SE researchers use or discuss financial incentives. So, we can reason on the extent of awareness for financial incentives and collect typical experimental designs for our interdisciplinary discussion on whether such incentives can be useful.
> RQ_2 *What forms of financial incentives are applied?*

From all experiments that employ some form of financial incentives, we elicited how exactly these are applied. So, we can understand potential benefits and limitations of financial incentives used in SE experiments.

Note that, due to our interdisciplinary analysis and terminology issues in SE (Schröter et al. 2017), experiments in our analysis include not only controlled, quasi, field, or simulation experiments; but also field and observational studies, which often resemble or actually are types of field experiments, particularly compared to other disciplines (e.g., lab-in-the-field experiments in experimental economics). So, such observational studies can exhibit similar properties in terms of motivating participants and using incentives. We involve such different types of studies to obtain a broader overview of the current state-of-the-art, and remark that typical setups for such SE studies are somewhat of a gray area considering the methods of other disciplines. For example, the differences between lab or field experiments and observational studies in SE are often blurry, since the environment and control opportunities of the lab and field are typically more alike than in other disciplines.

3.2 Search Strategy

Automated searches are problematic to conduct and replicate, due to technical issues of search engines (Kitchenham et al. 2015; Krüger et al. 2020; Shakeel et al. 2018); and our test runs resulted in many irrelevant results. For instance, the search string

```
"financial incentives" AND
"software engineering" AND
experiment
```

returned roughly 1,560 results on Google Scholar, which report on healthcare, machine learning, or motivators of software engineers—but rarely involve SE experiments. Identically, our searches on dblp and Scopus (cf. Section 2.4) returned no viable datasets. Therefore, we decided to conduct a manual search instead.

We aimed to cover a representative set of up-to-date best practices in SE experimentation. For this purpose, we decided to perform a manual search, covering the years 2020 and 2021 (to include recent publications that were available and officially published) of high-quality SE venues. Namely, we analyzed six conferences and five journals that involve empirical research:

Int. Conf. on Automated SE (ASE)
Int. Conf. on Evaluation and Assessment in SE (EASE)
Int. Symp. on Empirical SE and Measurement (ESEM)
Int. Conf. on Program Comprehension (ICPC)
Int. Conf. on SE (ICSE)
Eur. SE Conf./Symp. on the Foundations of SE (ESEC/FSE)
ACM Trans. on SE and Methodology
Empirical SE
IEEE Trans. on SE
Information and Software Technology
J. of Systems and Software

Note that this selection may have introduced bias, since we did consider two recent years and 11 venues only. However, if financial incentives in SE experimentation were an established concept, they should be used and reported appropriately in experiments recently published

at these high-quality venues. So, we argue that this sample of publications is sufficient to obtain an overview understanding of the state-of-the-art.

Furthermore, we acknowledge that the COVID-19 pandemic could have had an impact on how financial incentives were used in experiments with human participants in 2020 and 2021. Please note that it is highly unlikely that a full experiment can be conducted, analyzed, documented, reviewed, and published within one year; and many deadlines for the venues we analyzed are in the year before the publication. As a consequence, the COVID-19 pandemic should have had almost no impact on the publications from 2020. Our results (cf. Table 3) further indicate that, in 2021, the number of experiments dropped slightly and more were conducted online (49 in 2021 versus 56 in 2020), even though we analyzed a larger number of publications (1,303 in 2021 versus 981 in 2020). This may be due to the COVID-19 pandemic preventing laboratory sessions. However, there is no apparent difference regarding the types and forms of incentives used or the ratio of publications reporting to have involved incentives, which is why we argue that the COVID-19 pandemic does not threaten our results.

3.3 Selection Criteria

We defined four inclusion criteria (ICs) for any publication:

IC_1 Reports an experiment or observational study in which the tasks have a certain solution that allows to measure a participant's performance.
IC_2 Reports a study involving human participants.
IC_3 Has been published in the main proceedings of a conference or represents a full research article of a journal (e.g., excluding corrections, editor's notes, retractions, and reviewer acknowledgements).
IC_4 Is written in English.

Note that we only consider experiments in which the participants' behavior or performance is relevant (IC_1), not setups in which participants simply rate the quality of an artifact to serve as a baseline for a predictive model or perform tasks only to obtain a ground truth for testing models. We added this refinement on task performance during our second data analysis, when we found several publications employing such a setup (Karras et al. 2020; Nafi et al. 2020; Paltenghi and Pradel 2021). Finally, IC_3 also ensures that the selected publications have been peer-reviewed.

Moreover, we defined two exclusion criteria (ECs):

EC_1 We excluded publications that report on other empirical studies, such as surveys or interviews, which involve incentives only for improving participation, not as a reward for spending cognitive effort during a task.
EC_2 For conference papers only, we excluded those that report an experiment as a sub-part of their contribution, typically to evaluate a tool.

We employed EC_2 to provide a better overview of current best practices. Mainly, we expected that the more restricted space that is available for a tool evaluation in a typical conference paper leads to missing details (which was confirmed when we scanned a set of these). As a consequence, we would potentially obtain a skewed perception of the SE community; not because incentives are not used, but because of missing details in technical papers. We did not employ this EC for journals, since SE journals typically do not enforce or have more relaxed page restrictions. Also, if publications that are concerned solely with an experiment do not report on (financial) incentives, we would not expect technical publications that perform an experiment only for evaluating a tool or technique to do so.

3.4 Quality Assessment

A quality assessment in an SLR aims to capture the quality of the involved publications (Kitchenham et al. 2015). For us, such an assessment would only be important if we would intend to compare the results of the experiments to understand which results are more reliable. However, we are concerned with understanding how (financial) incentives are used in SE experiments, which in itself can represent a quality criterion since financial incentives can improve an experiment's validity. Consequently, we did not employ a quality assessment for the publications we identified during our SLR.

3.5 Data Extraction and Collection

For each selected publication, we extracted the relevant bibliographic data from dblp into a spreadsheet, which we used throughout our whole study to add, store, structure, and analyze data. To enable a sophisticated analysis regarding the *use of financial incentives*, we collected data on the experiments' *context* and on typical criteria used in experimental economics to assess these incentives' design, use, and quality—which connect to the three conditions *dominance*, *monotonicity*, and *salience* (cf. Section 2). In total, we further extracted the following data:

- The scope and goal of the study (*context*).
- The measurements used and whether statistical tests or effect sizes are reported (*context*).
- The experimental design (i.e., within/between subject), number and profile (e.g., students) of participants, as well as number of treatments and consequent participants (*context*).
- Whether incentives are described at all, and if so their value like a brain model or the concrete monetary amount (*use of financial incentives*).
- Whether a payoff function was used (with details on differences between treatments, tasks, designs) and how it was defined, or whether the incentives represent a show-up fee, completion fee, lottery, or any other form of payoff (*dominance*, *monotonicity*, *salience*).
- Whether and what fixed amount of time was allocated for a participant to perform their tasks (*dominance*).
- What the hourly wage of a participant would have been (i.e., comparing earned incentives to the amount of time allocated), and whether this wage is somewhat realistic for the country in which the experiment was performed as well as the participants (*dominance*).

We added this data to our spreadsheet to ensure traceability and allow us to perform our interdisciplinary analysis. Note that we did not find many details on most of these entries, which is why we cannot reliably analyze and compare them (e.g., whether the payment represents the hourly wage based in the allocated time and country). Still, we required this data to provide a detailed understanding for all authors.

3.6 Conduct

Collecting Bibliographies The first two authors of this article extracted the bibliographic data of all conferences and journals from dblp into spreadsheets, resulting in a list of 2,284 publications (cf. Fig. 1). Before our analyses (divided by years), we aimed to remove all publications that do not belong to the main track of a conference or are (one page) corrections in journals by considering the information of dblp and the number of pages. Namely, we

removed all conference papers with fewer than eight pages to discard, for instance, tool demonstrations, data showcases, or keynote abstracts. However, it was not always possible to clearly identify industry papers at conferences, since they can have a similar length to typical main-track papers and are sometimes insufficiently marked in the proceedings. As a consequence, the number of publications for each step in Fig. 1 may be a bit higher than the actual number of the official research publications.

Selecting Publications from 2020 For the 981 publications from 2020, the first and second author independently iterated through all publications and decided which to include based on our selection criteria (by assigning "yes," "no," or "maybe"). We compared the individual assessments to reason on the final decision of including or excluding a publication. Both authors agreed on 927 publications, while they disagreed on 54 (mostly, one author marked the publication with a "maybe," while the other stated a clear "yes" or "no"). Overall, we achieved a substantial inter-rater reliability, with a percentage agreement of 94.50 % and Cohen's κ of $\approx$0.7 (counting every "maybe" as a "no" for κ). We resolved disagreements by re-iterating over the respective publications and discussing the individual reasonings. This also led to refinements regarding our selection criteria (e.g., we adopted EC_2 so that it covers only conferences, but not journals). In the end, we considered 60 publications to be relevant for our SLR.

Then, the first two authors split the selected publications among each other and manually extracted the data for their subsets. For this purpose, they read each publication, focusing particularly on the abstract, introduction, methodology, and threats. Furthermore, they used search functionalities to ensure that they did not miss details, for instance, by searching for the term "incentive." When in doubt about the details of a publication, the other author cross-checked the corresponding publication. In the end, the two authors performed a cross-validation of the extracted data, investigating the other author's subset. Afterwards, the remaining authors analyzed whether the data was complete and sufficiently detailed for them to understand each experiment—leading to our refinement of IC_1 and the exclusion of four publications, resulting in a total of 56 publications. We remark that it is challenging to identify whether IC_1 applies until reading the details of a publication, which is why we also excluded some publications during the data analysis. During the validation, we added some details for individual publications, but did not find major errors. Mostly, we clarified some SE context on the experiments for the authors from experimental economics and behavioral psychology or corrected an entry (e.g., changing the number of participants).

Selecting Publications from 2021 After refining our selection criteria and obtaining a common understanding on the experiments, we continued with the 1,303 publications from 2021. For these, the second author employed the selection criteria alone. The first author validated a sample of 152 (11.67 %) publications, including all 56 marked as "yes" or "maybe" as well as 100 randomly sampled ones (overlap of four). Regarding this sample, both authors agreed on 145 publications and disagreed on seven. So, we achieved an almost perfect inter-rater reliability with a percentage agreement of 95.39 % and Cohen's κ of $\approx$0.87 (counting every "maybe" as a "no" for κ). We considered 51 publications as relevant and the second author extracted the corresponding data. The first author validated the whole dataset and refined some of the entries. We excluded two more publications due to our refinement of IC_1 based on discussions with all authors, leading to a total of 49 publications.

Triangulating and Analyzing Finally, we analyzed the data of all 105 remaining publications (56 from 2020, 49 from 2021). We remark that several publications involved multi-method study designs, for instance, combining surveys with a later experiment. In such cases, we extracted only the data on the experimental part. We performed our interdisciplinary analysis mainly in the form of repeated discussions, which started during the design of an actual

experiment (Krüger et al. 2022). After conducting our SLR, we inspected the results and compared the reporting and use of incentives to best practices and guidelines in experimental economics as well as behavioral psychology. For this purpose, we built on the expertise of the respective authors from each field as well as the established guidelines and related work that we summarized in Section 2. Specifically, the respective authors iterated through the data, took notes, and investigated different experimental designs based on the actual papers to understand how experiments are designed in SE and outcompared those to their fields. Then, we continuously discussed their impressions, the motivations of software engineers, and the context of SE experiments (e.g., considering open-source developers' motivations) to understand differences between the fields. Primarily, we discussed the general results of our SLR, which we present in Section 4. Building on more than 30 hours of discussions, individual analyses, and synthesis, we then derived and iteratively refined our guideline and recommendations in this article. At this point, we particularly considered established guidelines from experimental economics (cf. Section 2) and adapted these to the specifics of SE.

Informing the Design of Guidelines (G) and Recommendations (R) The data (cf. Section 3.5) we collected during the SLR (e.g., prevalence and type of financial incentives, background and number of participants) is based on criteria from experimental economics and psychology. This data is used to investigate whether an experiment correctly applied financial incentives and correctly disclosed the application of these incentives. Both aspects are important to facilitate replications of experiments, and thus to increase the quality of experiments. If our SLR indicated failures in correctly applying financial incentives and disclosing them, this justifies the need for defining precise recommendations customized for SE. Additionally, besides capturing whether (RQ_1) and what (RQ_2) financial incentives are used in SE experimentation, our SLR was also the basis to investigate whether financial incentives should be used in SE. Doing so enables us to not merely copy guidelines from other disciplines, but to develop *SE-specific* guidelines and recommendations. To achieve this, we first identified the relevant types of studies (i.e., IC_1, IC_2) that are related to those used in experimental economics and psychology. This was important, because some SE experiments did not involve measurements related to participants' performance. However, the prevalence of financial incentives is especially important for experiments where performance plays a role. For instance, this resulted in Q_1 in our guidelines.

We further extracted the selected data to inform our interdisciplinary analysis. Concretely, the measurements helped us distinguish again whether and what parts of an experiment were connected to performance or not (e.g., R_7). The experimental design and population are important because these can impact the design of payoff functions (e.g., R_1, R_2, R_5) and also reveal specialized populations that require different means. For example, open-source developers that work for free are somewhat known in experimental economics, but rarely studied. Reflecting on their motivations from the lens of psychology helped us understand how to translate these into financial incentives (e.g., Q_{4a}–Q_{4c}, R_7). The question what incentives are used and for which experimental designs is important for our analysis to understand how financial incentives could be employed altogether. Overall, the criteria we used to extract data for the SLR are based on studies from experimental economics and psychology, and should help us judge what properties to consider and adapt in what form. As a concrete example, we identified that some experiments relied on course credits, which are discouraged in experimental economics for their lack of comparability and replicability.

Identically, through discussing and analyzing our data, we obtained further general insights (cf. Section 4.3) that were important for our guidelines and recommendations. For instance, one findings was that SE experiments do not document the use of (financial) incentives well. Therefore, we stressed documentation as an important recommendation for SE experimenta-

tion to improve their replicability and comparability (e.g., R_{10}, R_{11}). Also, we discussed that various specific populations are participating in SE experiments. From the point of experimental economics and psychology, the smaller sample sizes caused concerns over the statistical validity and generalizability of results. Consequently, we also incorporated such specifics in our guidelines (e.g., Q_7, Q_8, R_3, R_9) and recommendations. Again, we argue that the general insights we obtained and discuss clearly support our goal of developing SE-specific guidelines and recommendations for using financial incentives during experiments.

4 Results and Discussion

In Section 2, we described financial incentives from the perspectives of experimental economics and behavioral psychology. Within this section, we describe and discuss the findings from our SLR based on these perspectives.

4.1 Results

We analyzed 105 publications. The studies reported in these publications aim to address a variety of goals in different scopes, for example, to understand the impact of being watched on developers' performance during code reviews (Behroozi et al. 2020). In Table 3 in the Appendix, we provide an overview of core properties of each study, namely the research method, study design, participants, and incentives. Note that the studies have been conducted with participants from a variety of countries, such as the USA, Canada, Germany, UK, Chile, or China—indicating that we cover a broad sample.

Overall, 76 publications cover experiments (e.g., online, quasi, controlled), 26 cover observational studies (e.g., fMRI studies, session recordings), and three cover both (cf. Table 3). Of the 79 publications reporting an experiment, 38 involve between-subject, 31 within-subject, and 10 hybrid (i.e., both or multiple experiments) designs. We remark that this distinction is infeasible for observational studies, since developers are exposed to the same treatment only once to explore patterns in their behavior. A majority of 76 studies involved (44 solely) students, 48 developers, 10 researchers, and four non-computer science participants. Note that the publications often report only on involving developers, without specifying the developers' concrete background. We summarize these studies as well as those that mention, for instance, industrial backgrounds, professionals, or API developers, under this term. Finally, the number of participants in the studies varies substantially (i.e., 6–907 for experiments, 4–249 for observational studies), and in 22 cases the exact number of discarded observations are explicitly described. We provide the number of valid observations in parentheses in Table 3, which are typically smaller because participants did not finish the study (e.g., in online settings (Spadini et al. 2020)). Overall, we argue that these 105 publications span a variety of experimental designs in SE and are a representative sample of the best efforts in current SE experimentation. Consequently, they were a feasible foundation for our analysis and discussion of using financial incentives in SE experimentation.

Unfortunately, the details on incentives are often vague, particularly regarding when they have been paid. Thus, we assumed that incentives were typically awarded for (valid) completions, and deviated from this strategy only if the publication hinted at, or explicitly specified, a different one. We mark forms of incentives for which we have been unsure with "(?)" in Table 3.

Table 2 Summary of how incentives have been payed out across the 105 publications

incentives format	publications		incentives					
	#	%		vouchers				
			money	money	meal	donation	credits	other
none	58	55.24	—	—	—	—	—	—
completion fee	28	26.67	16	2	2	2	11	2
with quality check	3	2.86	1	—	—	1	1	—
min.–max costs			125€—$12,540	2,150€—$2,950	—	$425—$730	—	—
show-up fee	12	11.43	4	2	—	—	6	—
min.–max. costs			$510—740€	$375—$849	—	—	—	—
wage/contract	4	3.81	4	—	—	—	—	—
min.–max. costs			$630—$4,180+	—	—	—	—	
piece rate	4	3.81	—	—	—	—	4	—
tournament	1	0.95	*<rewards mentioned, but not explained>*					
lottery	1	0.95	—	1	—	—	—	—
min.–max. costs			—	$30	—	—	—	
total #	—	—	25	5	2	3	22	2
% publications	—	—	23.81	4.76	1.90	2.86	20.95	1.90

We indicate the minimum to maximum costs of the experiments (under 1:1 exchange rates), as far as we can estimate these costs from the publications. Please note that the numbers and percentages do not add up to 105 (100%), since some experiments employed multiple incentives for different or even the same populations

In Table 2, we synthesize an overview of the incentives used within the studies. We can see that a majority of 58 (55.24 %) publications does not report on using incentives. Note that one publication (Wyrich et al. 2021) states that the participating students were required to participate in a study to fulfill the university's curriculum, which we did not consider as an incentive on its own (i.e., without further incentives for their efforts). Out of the remaining publications, most used direct money payments as incentives (25, 23.81 %) followed by course credits (22, 20.95 %). However, almost all of the money payments are connected to fixed payments, but not payoff functions: In 28 (26.67 %) publications, the use of fixed incentives awarded after completing the experiment are mentioned, including fixed amounts of money for each participant, donations to charities, and non-financial incentives (e.g., images of brain scans from fMRI studies, course credits for students). Another 12 (11.43 %) publications (seem) to refer to a show-up fee in the form of course credits or a fixed amount of money. Four (3.81 %) publications mention to have paid participants hourly wages or provided them with contracts, paying them directly with money or with gift cards. Interesting is one study that investigates how different payments (hourly wages versus fixed contracts) impact freelances (Jørgensen and Grov 2021), indicating that fixed contracts led to higher costs. One (0.95 %) publication (Bai et al. 2020) used a lottery among the participants to distribute a voucher.

Some publications indicate that more advanced incentivisations or even payoff functions have been used, but the details are still lacking. Four (3.81 %) publications (seem to) have used course credits as a piece rate to reward a participant's performance. Three (2.86 %) publications indicate that a quality check was performed before paying out a completion fee, which makes the payoff somewhat performance-based since only the correct solution yields a payoff (cf. Table 1). Finally, one (0.95 %) publication (Shargabi et al. 2020) reports on rewarding only the best-performing participants (i.e., winners-take-all tournament), but not what that reward constitutes.

Lastly, we can see in Table 2 that our cost estimates span from \$30 to \$12, 540. Unfortunately, the information on these costs is often vague or incomplete, too. For instance, one publication (Jørgensen and Grov 2021) mentions contracts, but only lists the smallest and highest paid contract. The publication with the "most expensive" incentives (Endres et al. 2021a) reports to have paid participants for contributing to multiple sessions (\$20 for each out of a maximum of 11 sessions). Unfortunately, it is unclear how many participants joined each session, so our estimate of \$12, 540 is an upper bound. We remark that a majority of 18 (out of 25 specifying a monetary amount) of the financial incentives resulted in costs (way) below \$1, 500, with average costs of \$1, 715.16 based on our estimates.

4.2 RQ_1 & RQ_2:The Use of Financial Incentives

We can see in Table 2 that roughly 44.76 % (47 of 105) of the publications report on some form of incentivisation. Of these 47, 30 (63.83 %) used financial incentives (i.e., indicated by a concrete monetary value or mentioning payments) and 37 (78.72 %) used fixed completion or show-up fees (involving non-financial incentives), which are similar to incentives used to improve participation in surveys. Since many experiments have been conducted with students, course credits have often been used as a replacement for actual financial incentives. This aligns to our personal perception of the research conducted in empirical SE, and underpins that our research tackles an important gap in SE experimentation.

RQ_1: To What Extent are Financial Incentives Used?
Financial incentives are somewhat used in SE experimentation, covering 30 of 105 (28.57 %) publications we analyzed. They constitute the primary form of incentivistation in our dataset (30 of 47 publications, 63.83 %).

Only 12 of the 47 publications hint at more elaborate strategies for incentivizing participants, such as payoff functions. Four studies have used, or at least indicated to have done so, course credits as a piece rate (Paulweber et al. 2021a, b; Taipalus 2020; Czepa and Zdun 2020). For instance, Paulweber et al. (2021a, b) rewarded students with up to six course credits for correct answers during experiments. However, the details of these studies (e.g., how the credits were assigned, how much they benefited the participants) are unclear. Also, course credits do not allow to replicate such experiments easily, a problem financial incentives can help to mitigate (cf. Section 4.4). Another four studies have paid their participants contracts and hourly wages to compensate for their time (Jørgensen et al. 2021; Jørgensen and Grov 2021; Liu et al. 2021; Aghayi et al. 2021). Two publications (Sayagh et al. 2020; Braz et al. 2021) indicate that performance-based financial incentives were used by quality checking the submitted solutions before paying a completion fee. For instance, Sayagh et al. (2020) paid five freelancers among their participants only after checking the quality of the solutions (students received course credits), so only a good enough performance allowed them to obtain the reward. Still, it is unclear to what extent the freelancers were informed in advance. Shargabi et al. (2020) used a winners-take-all-tournament, awarding the best performing students with an unspecified incentive. The incentives in all of these experiments are rather simplistic performance-based payoffs, and more elaborate payoff functions as used in experimental economics have apparently not been used. Overall, only five (4.76 %) of these studies report on using financial incentives, and thus may fulfill the properties of a payoff function (cf. Section 2).

RQ_2: What Forms of Financial Incentives are Applied?
While financial incentives are used, they are mainly applied as a mechanism to motivate participation (e.g., completion fee). More advanced techniques that (may) cover all properties of payoff functions to motivate behavior during a task and increase validity are rarely employed (5 of 105, 4.76 %).

4.3 General Insights

When analyzing the individual experiments, particularly the authors from other disciplines raised several discussion points to move towards our guidelines and recommendations. In the following, we summarize the four major points as general insights on incentives in SE experimentation.

Reporting Incentives We already noted that it was sometimes problematic to understand how incentives were used during an experiment. For instance, we re-checked various publications numerous times trying to obtain a better understanding regarding whether incentives were paid on completion or as show-up fees. In fact, many publications simply stated to have used some form of incentivisation without providing any details (e.g., students have been graded somehow). This finding highlights the need for improving our understanding on how and when to employ incentives in SE experimentation, and what details to report. For example, a majority of 58 publications simply does not mention incentives, but they also do not specify that they have not been used. Similarly, few experiments report on the reasoning for using

a particular form of incentives. For instance, LaToza et al. (2020) indicate to rely on a fixed show-up fee (it may have also been a completion fee) to not bias their participants. Such details are important to allow researchers to understand, evaluate, and particularly replicate previous experiments. In experimental economics, it has become standard to rigorously describe the incentivisation. The information typically includes the average payoff of the participants (consisting of show-up fee and performance-based payments), whether all decisions were relevant for the payoff, any influence of chance, and the average duration of the experiment. Also, it has to be clear that the applied incentives were identical for all participants, unless incentives were a treatment variable.

Population Sizes One issue raised in our discussions are the relatively small population sizes of some experiments we reviewed. Even though there are statistical tests that can be used (e.g., Fisher's exact test) and the studies are still meaningful to understand developers' behavior, it remains an open issue to what extent the findings can be transferred to the general population. Considering also that many populations involve convenience samples (e.g., students of one particular course of one university), the population sizes of the experiments are an important concern that should be considered carefully. However, discussing the (mis-)use of statistical tests and significance is out of scope for our work and we refer to existing research (Wasserstein and Lazar 2016; Wasserstein et al. 2019; Amrhein et al. 2019; Baker 2016). Still, this issue highlights the potential for improving participation and the external validity, for instance, by using financial incentives.

Wages, Realism, and Participants Incentives in experimental economics are used to mimic realism and compensate participants' time by resembling the incentive structure of the real world—typically paying for participants' time/effort oriented towards real-world wages. When reflecting on and discussing these intentions with respect to the developers involved in the studies we reviewed, we also raised the issue of (the subset of) open-source developers who contribute to their projects without receiving money. For them, financial incentives may not actually mimic the real world (e.g., if they are unpaid contributors). However, financial incentives can also help to mimic factors (open-source) developers may otherwise not be aware of and for which no other incentive is viable in an experiment (cf. Section 5), such as a loss of reputation in the community.

Similarly, we found publications that employed observational studies or experiments during action-research with industry. In such cases, incentives may also not be useful or even applicable. As aforementioned, for these cases the incentivisation stems from the work itself and additional performance-based incentives should actually be avoided to prevent biases. However, one study also explicitly mentions that companies were compensated for their developers participating in a study by the researchers paying the participants' wages for the time of the conduct (Jørgensen et al. 2021). These observations highlight that (financial) incentives are not a silver bullet for SE experimentation, but they should be considered and their non-use reported as well as explained.

Laws, Compliance, and Ethics The laws and ethics around financial incentives are of particular importance when designing an experiment, and can vary depending on the location at which a study is conducted. For instance, in our SLR, researchers mention that they were not allowed to pay students for participating in experiments, whereas bonus course credits were allowed (Baldassarre et al. 2021). As a result, financial incentives may not be usable in experiments conducted in the respective countries or universities—but please note that we could not identify any references that support the argument that students could not be paid legally. This situation as well as the different populations involved in experiments (e.g., open-source developers versus students) raise ethical issues. Namely, it is questionable to conduct an experiment with different populations and rewarding only some of the participants (with

money) or to change the average payoff. For instance, some experiments reported to have compensated students with course credits or not at all, but other participants received (inherently more) money (Sayagh et al. 2020; Uddin et al. 2020; Danilova et al. 2021; Uddin et al. 2021). Moreover, research (Różyńska 2022) underpins that it is not only fair to compensate participants with financial incentives for their efforts, but it is a moral obligation to do so in a fair and ethical way without discrimination or exploitation. Therefore, when designing the incentives for an experiment, researchers have to keep legal and ethical implications in mind (cf. Section 6). One particular concern in this direction that researchers have to be aware of are compliance issues—depending on whether professionals participate during their work or free time. Some companies may forbid their employees to receive external money, which means that financial incentives cannot be used. However, in other cases, using financial incentives may even be required to reassure a company and facilitate industrial collaboration, thereby lowering the risks of failure and increasing the value of the study (Sjøberg et al. 2007).

4.4 SE and Payoff Functions

Our SLR clearly shows that advanced financial incentives are sparsely used (or reported) in SE experimentation. Precisely, if incentives are used, they rarely combine monetary, task-related, and performance-based incentivisation. In fact, only the four studies paying participants hourly wages or contracts (Jørgensen et al. 2021; Jørgensen and Grov 2021; Liu et al. 2021; Aghayi et al. 2021) and the one study assessing the freelancer's solutions (Sayagh et al. 2020) involve all of the concepts to some extent. Still, whether and what payoff functions have been used is not reported in the publications. In the following, we discuss how payoff functions connect to SE experimentation based on the results of our SLR.

Dominance Condition Payoff functions should mimic real-world incentives so that the dominance condition holds (i.e., incentives must be strong enough to out-power other aspects that can motivate participants' behavior). For instance, in their functional Near Infrared Spectroscopy study, Endres et al. (2021b) paid their participants $40 for 2.5 hours, which is equivalent to $16 per hour. Based on six reports, `indeed.com`[6] indicates an average hourly wage of around $21 for an intern developer in Michigan, USA. So, the payment in this study seems to properly mimic the real world (also depending on taxes), which is why we could assume that the dominance condition holds as long as no other aspects were involved (e.g., students participating as a mandatory requirement in the curriculum).

We cannot assess whether the dominance condition holds for online studies we found during our SLR, since the participants can be from any country and average hourly wages vary. Moreover, wages also vary between states or cities even within the same country (e.g., USA). We found 12 (11.54 %) studies that were conducted online (e.g., using customized online experiment or crowdsourcing platforms, such as `MTurk`,[7] `Freelancer.com`[8]). Similarly, another 14 (13.33 %) studies involved participants from multiple countries. For all of these studies it is challenging to calibrate the payments in a payoff function, considering that they must be adopted to local payment structures to fulfill the dominance condition—while also paying attention to local regulations (e.g., not allowed to pay students) and consequent ethical issues (e.g., paying participants different incentives).

[6] https://www.indeed.com/career/software-development-intern/salaries/MI?from=top_sb (September 27, 2022)

[7] https://www.mturk.com/

[8] https://www.freelancer.com/

These issues can be tackled in various ways. For instance, conducting laboratory sessions in-person in one city mitigates these problems, while potentially limiting the external validity. Another way is to limit the countries the participants can come from (e.g., using `Prolific`,[9] an on-demand platform that connects researchers to participants) so that the hourly wages in all considered countries are close to each other. We want to remark again that payoff functions are not feasible for all types of experiments, and thus these design decisions do not impact all studies (i.e., if the tasks are not responsive to performance).

Monotonicity Condition Payoff functions should align with the monotonicity condition, meaning that participants prefer more of the incentive over less of it. A majority of the studies using (financial) incentives we investigated employed a show-up or completion fee, with a completion fee representing a flat payoff function. However, a flat payoff function does not incentivize effort (cf. Table 1), and thus may not be appropriate for studies in which effort is required to perform well. For instance, Bai et al. (2020) used a lottery to distribute an Amazon gift card among their participants. Unfortunately, rewarding participants with gift cards does not impact participation rates in the same way as money (Becker et al. 2019; Veen et al. 2016) and can also induce varying levels of effort, since people value gift cards of the same monetary value differently (Gunasti and Baskin 2018). As examples, if participants do not have access to the company, are not its customers, or even purposefully avoid it, we cannot ensure that the monotonicity condition holds. We found few studies that indicate to have used payoff functions for which the monotonicity condition holds, like those by Paulweber et al. (2021a, b).

Salience Condition Unfortunately, even for the studies for which monotonicity holds, it seems that the salience condition may be violated (i.e., it is unclear what incentives are paid for what performance). The descriptions of the payoff functions are unspecific, hidden, and scattered in the text; leaving the functions' interpretation to the reader. Consequently, it is unclear to the reader and has likely been unclear to the participants (since it is not specified in the publications that they have been informed about the mapping) how their performance maps to the incentives. From the perspective of experimental economics, such details must be reported, also to ensure that the study can be replicated and does not face additional threats to its validity.

Financial Incentives versus Course Credits Since 20 (19.05 %) studies in our SLR relied on course credits as incentives, such credits are motivated as a compensation for students in SE guidelines (Carver et al. 2010), and some researchers apparently are not allowed to or cannot afford to pay large amounts of money, the question arises *whether we should have payoff functions for course credits?* Unfortunately, while course credits are a viable compensation, they are not comparable to financial incentives, and can pose a threat to the replicability of an experiment. When the incentives are monetary payments, we can compare between participants (e.g., different countries, universities) and consequent experiments, since information about currency conversions and the average wages in countries as well as regions are available. Such information also facilitates comparing across different countries and replications.

As a concrete example, we could argue that we can convert course credits in a similar fashion as money, for instance, 1 ECTS at an EU university represents two credits in the UK. However, such transformations are not always possible, and it is usually unclear how a specific number of credits contributes to a student's overall grade. Additionally, how students are graded may vary even within the same university or country (e.g., in Turkey, some universities award a letter grade as in the USA, while some universities award a grade in the range

[9] https://www.prolific.co/

1–10 with 10 being the highest grade). Finally, we would need to know the entire context of the course. For instance, if an experiment offers students four credits for completing the tasks, we do not know whether there is an easier way for the students to obtain the same number of credits. Consequently, we cannot understand to what extent four credits actually motivate the students; particularly, since the student's goal may be to only pass that course. In such a case, the dominance condition may not be satisfied (i.e., other factors may motivate students' behavior during the experiment, such as finding the task interesting). These issues are besides the limiting fact that course credits allow to conduct experiments with students only, excluding and challenging comparisons to other participants for which other incentives must be employed. To summarize, course credits are not an equivalent replacement for financial incentives, should be used very careful when deciding about incentivizing participants, and payoff functions that build on course credits can basically not be properly replicated.

5 Guideline for Using Financial Incentives

A payoff function is a mapping between the set of choices participants can make in an experiment to their financial payoffs. This makes it possible to calculate the financial payoff for every participant based on the choices they made in the experiment. Behind this principle lies the idea that the production or consumption of every good or service has a monetary equivalent—a willingness to pay for this good or service. Inducing these monetary values into the experiment is the purpose of a payoff function. Experimental economics provides methodological critique on a superficial implementation of payoff functions (Harrison 1992) and the response to this critique (Merlo and Schotter 1992; Smith 1994). Still, for the majority of all experiments it is sufficient to consider the three major conditions for payments that we described in Section 2.1 and discussed in relation to our SLR results in Section 4.4: dominance, monotonicity, and salience. Following these principles enables researchers to implement a payoff function even without having a precise game-theoretical model underlying the experiment, which is typically the case in experimental economics.

Most importantly, the incentive structure should be similar to the real-world case, while covering the participants' opportunity costs (e.g., hourly net wages) of participating in the experiment. More precisely, experimental economics typically has a well-defined understanding of such costs (e.g., based on existing studies and measured expenses), which is missing and challenging to elicit in SE. So, designing payoff functions is arguably also challenging considering the complexity of SE and the cognitive processes involved.

As a first step, SE researchers have to identify whether using a payoff function would benefit their experiment. Building on our insights from the SLR (cf. Section 4) as well as guidelines on incentives from experimental economics and behavioral psychology (Schram and Ule 2019; Weimann and Brosig-Koch 2019), we derived the flowchart we present in Fig. 2. Researchers can apply the flowchart to assess whether a payoff function can be applied. We now discuss the individual steps of the flowchart and exemplify its use in Section 7.

Our guideline involves 11 questions (Q) that researchers should consider when designing their experiments in SE. We adapted Q_{1-3} for SE based on research in experimental economics. Q_4 and its sub-questions (Q_{a-c}) build on research from experimental economics and psychology, which we constructed primarily due to the specifics of unpaid open-source development we observed in our SLR. Q_5 and Q_6 are driven by psychological considerations. Q_7 and Q_8 are practical as well as ethical considerations that stem directly from observations in our SLR. For instance, Q_8 is concerned with the legality of payments to students, which

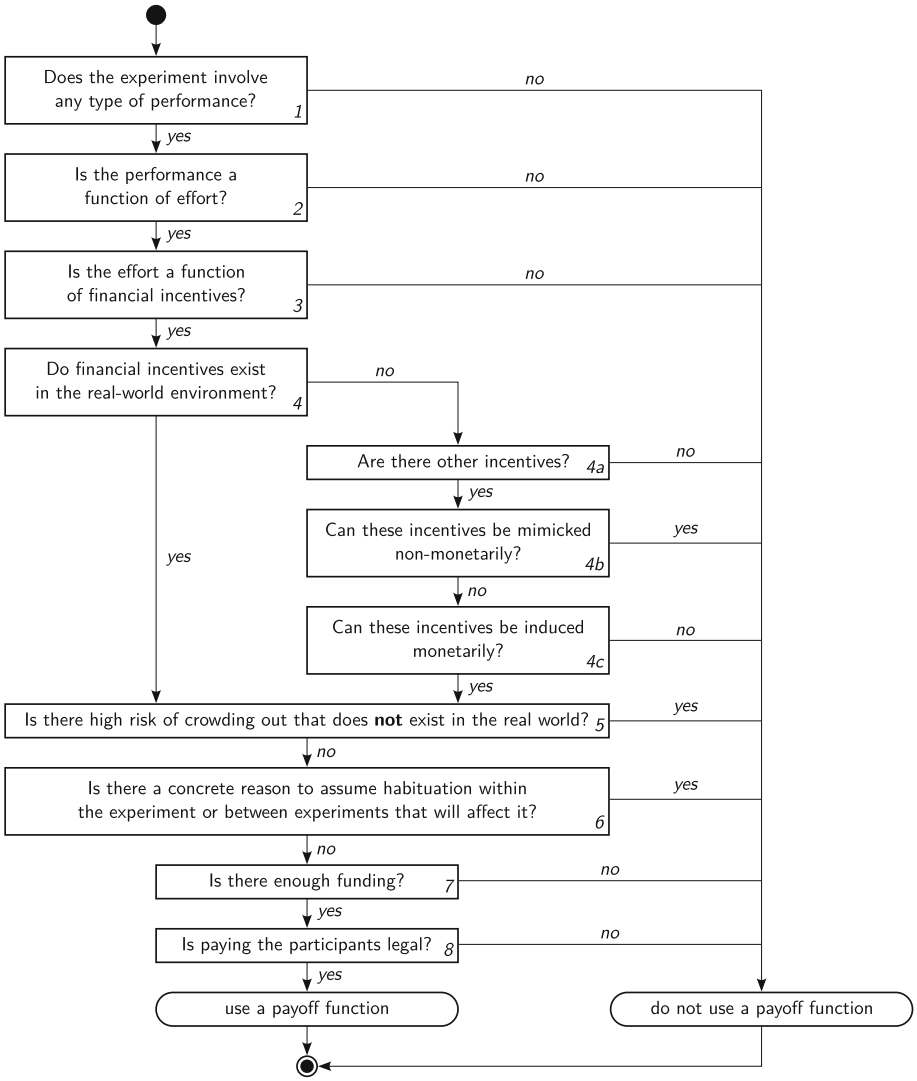

Fig. 2 Guideline for deciding whether to use financial incentives in an SE experiment

was an issue reported in one experiment (Baldassarre et al. 2021); which is not an issue in experimental economics and thus not covered in their guidelines. We compared our questions against sample experiments and the data from our SLR to ensure the applicability of these questions for designing SE experiments:

Q_1 *Does the experiment involve any type of performance?*
A payoff function (i.e., not only a flat show-up or completion fee to motivate general participation) is only useful if the experiment is impacted by the participants' performance (e.g., solving programming tasks). If this is not the case (e.g., collecting opinions), there is no need for a payoff function.
Q_2 *Is the performance a function of effort?*

When performance is relevant, the researchers have to identify whether the performance is impacted by the effort their participants spend (e.g., performing program comprehension). If the performance is not impacted by effort (e.g., having someone work with a language they have no idea about), a payoff function is not useful.

Q_3 *Is the effort a function of financial incentives?*

Next, the researchers have to understand whether the effort can be impacted by financial incentives. If so (e.g., motivating more concentration during code reviews), a payoff function may be useful, otherwise (e.g., comparing reactions to a video) it is obviously not.

Q_4 *Do financial incentives exist in the real-world environment?*

Up to this point, it has become clear that a payoff function could be used. Now, the researchers have to identify whether there are financial incentives in the real-world setting, too. If this is the case, a payoff function can be used; and if not, a payoff function may still be applicable, but three more questions become relevant:

> Q_{4a} *Are there other incentives?*
> In the simplest case, there are no other incentives that can be mimicked with financial ones, meaning that a payoff function is not useful. Otherwise (e.g., reputation as incentive), a payoff function can be useful to replace non-financial incentives, too.
> Q_{4b} *Can these incentives be mimicked non-monetarily?*
> Consequently, the next question is whether the real-world non-financial incentives can be mimicked without money (i.e., to stay closer to the real world). If this is not the case, a payoff function is a useful option.
> Q_{4c} *Can these incentives be induced monetarily?*
> Finally, to substitute a non-financial incentive with a payoff function, it must actually be possible to replace the former with the latter. This may not always be possible (e.g., intrinsic motivation of working on own project), but quite often a payoff function is useful.

Q_5 *Is there high risk of crowding out that does* ***NOT*** *exist in the real world?*

Crowding out (cf. Section 2.3) means that financial incentives could replace intrinsic motivation, making it harder to assess their impact on the participants' performance—thus, threatening the validity of an experiment.As a consequence, if a payoff function has the risk of crowding out, the researchers should only use one if the same risk exists in the real-world setting (e.g., open-source developers receiving bug bounties).

Q_6 *Is there a concrete reason to assume habituation within the experiment or between experiments that will affect it?*

Habituation (cf. Section 2.3) has not been confirmed in experimental economics. Still, researchers should consider whether habituation effects (i.e., participants getting used to the financial incentive, reducing the motivation caused) could occur within their experiment or family of experiments. If this is not the case, a payoff function is definitely useful for the experiment, but it should not be used otherwise.

Q_7 *Is there enough funding?*

Depending on the size, type, and population of an experiment, it is easily possible that there is simply not enough money to pay for financial incentives. This is a completely acceptable reason not to use a payoff function, but should be reported.

Q_8 *Is paying the participants legal?*

Researchers have to check whether financial incentives are legally allowed in their experimental setup, and if not a payoff function should obviously not be used.

Note that our guideline moves from foundational decisions whether a payoff function would be useful down to the more practical decision whether it can actually be used. We recommend researchers to check in that order, and to report for what reason a payoff function is not useful (i.e., Q_1-Q_6) or whether there are practical limitations that cannot be overcome (i.e., Q_7, Q_8). Reporting this information improves the confidence, trust, validity, comprehensibility, comparability, and replicability of SE experiments.

6 Designing Financial Incentives

Reflecting on our findings, we argue that financial incentives would have mostly positive effects on the validity of several SE experiments. That is, they could improve sample sizes, mitigate selection bias, motivate the desired behavior, improve replicability, and allow to study additional properties of SE. Our guidelines in Section 5 help researchers decide whether to use a payoff function. In this section, we discuss 11 recommendations (R) on how to design such a function, which we also exemplify in Section 7.

These recommendations are based on our interdisciplinary analysis that was driven by the results of our SLR. Notably, we found that a majority (55.24 %) of the publications do not report on using (financial) incentives, which directly impedes replicability and leads to one of our recommendations (R_{10}). The SLR results further imply that there are SE researchers who are sensitive to the role of incentives as they apply performance-dependent incentives, such as course credits. Yet, this approach can be improved in different ways to increase the validity and replicability of an experiment. Likewise, the SLR results indicate that a substantial share of researchers have funding, yet it is mostly used for flat payments (e.g., show-up fees) even though performance plays a role in those experiments. Other experiments rely entirely on the intrinsic motivation of the participants. Thus, our findings imply that SE researchers should investigate more thoroughly whether the incentives are aligned with performance (R_6) and question the role of intrinsic motivation in their experiments (R_7). This implies that researchers should also explain the decision to (not) include financial incentives (R_{11}). Further, the SLR indicates that SE experiments are conducted in different countries. These different countries have varying levels of income, which has to be accounted for. Yet, our SLR indicates that it is impossible to compare the relation of payoffs to the local hourly wages. This contributes to our recommendations (R_6), indicating that payoffs should be noticeable for participants and therefore may vary dependent on the country of the experiment or the participants background. In total, the SLR indicates a need for improvements and leads the way to SE-specific recommendations on how to apply financial incentives in experiments, which we present in the following.

Strategy for Incentivisation Considering payoff functions, experimenters have to figure out how much money should be paid for what action. The most important construct for SE in this regard is the effort a participant is willing to put into solving the experiment's tasks. Precisely, the goal should be to keep the effort of the participants at realistic levels throughout the experiment. Thus, a realistic model of the studied situation is already an important precondition for a feasible incentivisation. Only if the experiment mimics realistic requirements, such as penalizing the misidentification of code as buggy, the payoff function can motivate participants to behave as in the real world—and allows to study time pressure or cost-benefit assessments (e.g., spending time on verifying actual bugs or finding as many as possible).

While all of the previous properties are performance-based and task-related (cf. Table 1), the characteristics of the participants themselves are equally important, since payoff functions work differently for different groups of participants. Most obviously, if highly experienced developers from industry with a potentially high salary participate, offering them the same amount of money as students may induce much less effort. This issue can be addressed by employing suitable opportunity costs, but is arguably harder to define in SE, considering the involvement of student, industrial, and open-source developers—each having a different background and motivation. Although it is likely not possible to elicit precise opportunity costs for each of the groups, it is possible to apply a sophisticated estimation based on the experience of the researcher. Similarly, participants with different backgrounds may respond differently to certain aspects of a payoff function. For instance, students may try to avoid penalties caused by misidentifying code as buggy much more than practitioners. However, if the goal is to actually compare such groups directly to each other (i.e., between-subject design), the same incentivisation should be used to avoid biases and ethical concerns (cf. Section 4.3).

Strategy for Incentivisation

R_1 Capture the real-world situation underlying the experimental design with respect to how task-related properties (e.g., rewarding or penalizing actions) and participant-related characteristics (e.g., defining feasible opportunity costs) relate to financial incentives of the tasks.
R_2 Conduct pilot studies (e.g., analyzing incentivisation tools used in practice, conducting exploratory experiments) to tune the financial incentives to the real-world situation.
R_3 Define financial incentives based on the experimental design, for instance, payoff functions may vary between tasks (i.e., within-subject), but should be constant for participants that are compared to each other (i.e., between-subject).

Time-Pressure The time required to solve a task is typically the most costly, and thus important, aspect (besides correctly solving the task) in SE practice. Consequently, this aspect should be mapped to corresponding experiments, which is typically done by measuring the time used or enforcing a time limit. Still, to induce time-pressure, it is also possible to use financial incentives (e.g., penalizing the time needed), since it is nearly impossible to create realistic time-pressure without performance-based incentives. Specifically, participants could simply put as much time and effort in solving a task as they want—depending on their intrinsic motivation (see next paragraph). However, in SE practice, many external drivers are relevant to developers, such as keeping a release deadline (i.e., time-pressure). As a concrete example, paying off in relation to the number of correctly identified bugs, while reducing the payoff over time, can realistically mimic time-pressure. Still, in an experimental setting, participants may simply identify much more code as buggy to obtain a high payoff, which is why penalties can be key (in this example, falsely identifying code as buggy should be penalized).

Time-Pressure

R_4 Reflect on how different incentivized actions impact each other to define suitable countermeasures.
R_5 Consider penalizing the time needed for a task as one of the most important aspects in SE practice.

Financial Incentives versus Intrinsic Motivation Even in experimental economics, the exact heights of financial incentives are debated between two extremes: payoffs do not matter versus sufficiently high incentives can trigger almost any level of effort (Weimann and Brosig-Koch 2019). Considering all evidence, financial incentives seem useful and can help to manipulate participants' effort, but it can be challenging to find the most suitable payoff function. Most experiments use rather small financial incentives (e.g., relating to typical hourly wages of students) to limit the costs. As a general rule, payoffs should be high enough to be noticeable for the participants (Weimann and Brosig-Koch 2019). So, it may not be necessary to compensate industrial developers fully based on their hourly wages, as long as the financial incentives are still valuable and help to trigger the desired level of effort.

Other forms of "extrinsic" incentives are social norms or a social status, but intrinsic motivation is an important driver, too. Research suggests that intrinsic motivation and extrinsic incentives are most often compensatory (Cerasoli et al. 2014; Locke and Schattke 2019). That is, if a participant lacks intrinsic motivation, their effort can be increased with extrinsic incentives. Identically, if a participant already has a high intrinsic motivation, paying financial incentives will also likely increase their effort—but to a lesser extent. For instance, many open-source developers contribute to their projects even without pay and sometimes with more effort than industrial developers. We imply that additionally incentivizing intrinsically highly motivated participants financially, such as open-source developers, will have negligible impact. In contrast, incentivizing intrinsically less motivated participants financially will have a substantial impact on their effort. Thus, controlling for the intrinsic motivation (i.e., measuring and analyzing it) may be an important contribution for explaining certain results. If financial incentives and intrinsic motivation are not in line or conflicting with regard to a specific task, they can lead to conflicting results and low validity.

Financial Incentives versus Intrinsic Motivation

R_6 Design financial incentives that are noticeable and aligned with the desired performance.
R_7 Consider the importance of intrinsic motivation and whether it varies between different groups (and if relevant, control it).

Project Status and Development Methodologies Interesting aspects of SE that we have to consider are the project status and development methodology, depending on which developers may vary the efforts they spend in real-world projects. For instance, in the beginning of waterfall-like methodologies, developers are only concerned with eliciting requirements, which is a continuous process in agile methodologies. Similarly, developers may start with creating more code in the beginning of a project, but over time fixing smaller bugs and improving the performance of the code becomes more important. So, when defining financial incentives, experimenters should also consider which project phases, development methodologies, and consequent technologies are relevant. Concretely, conducting an experiment on eliciting requirements or fixing bugs in waterfall-like and agile methodologies may require different financial incentivisation to account for the iterative versus continuous processes employed in the respective methodology. For example, finding a bug at the beginning of a waterfall-like methodology may be mapped to a smaller payment than in later phases.

Project Status and Development Methodologies

R_8 Assess whether different project phases or development methodologies may imply different efforts to participants.

Costs and Benefits Financial incentives promise several benefits regarding the number of participants, their motivation and behavior while performing their tasks, and the possibility to study their characteristics, for instance, regarding risk-taking. Still, financial incentives can be expensive, for example, because the opportunity costs for industrial developers are high or a large number of participants is needed. Moreover, as we discussed, not all experiments in SE can benefit from financial incentives (besides completion/show-up fees to increase the number of participants). To limit the costs of experiments and avoid that research funds are spent unnecessarily, experimenters should discuss the use of financial incentives early on. If financial incentives are a helpful means, we hope that this article provides help in designing them and reasoning to funding agencies why they are included in a project budget.

Note that the general impact of financial incentives on the experiment's quality should be positive. Particularly, adapting real-world incentives in SE experiments increases the external validity of the experiments. Unfortunately, there is the problem that potentially only some researchers with the required funding can afford using financial incentives. Yet, this is not a new phenomenon, since a lot of experiments require costly material that is not easily affordable (e.g., considering fMRI studies). Still, the consequence is that a higher validity may induce more costs, but we remark again (cf. Q_7 in Section 5) that not having the money for financial incentives is a completely valid reason not to use them—and it will not invalidate the results.

Costs and Benefits

R_9 Evaluate the costs and benefits of financial incentives against each other when designing an experiment.

Reporting Finally, we already mentioned the problems we had with eliciting the relevant data on incentives from current SE experiments. Therefore, we emphasize the importance of our last two recommendations to improve the comprehension and enable replications of prior experiments. Since incentives can motivate or change participants' behavior, they are key to experiments involving human participants. We recommend to document them in more detail, potentially within a dedicated section, to allow others to understand, evaluate, compare, and replicate an experiment. Precisely, we listed data we aimed to extract and that is relevant in Section 3, important design decisions in Section 5, and aspects to consider in this section. At least this information should be reported on the experimental design. We recommend to state explicitly if (financial) incentives have not been used. Moreover, either strategy, but particular non-incentivisation, should be explained and reasoned about to allow others to understand why this particular strategy and payoff function have been used; particularly for the sake of replications and comparisons. Lastly, if no funding was available for incentivizing an experiment (Q_7), we recommend to report that participants were not incentivized, what means were taken to obtain such funding, and to still report whether there is a real-world incentive that thus was not mimicked. This facilitates replications and can improve the trust that the experiment was designed rigorously.

Reporting

R_{10} Report all details on how participants were (financially) incentivized, or explicitly state that they were not.
R_{11} Reason on the design of financial incentives (e.g., pilot studies) or why they are infeasible (e.g., why they may introduce biases).

7 Exemplifying Payoff Functions in SE

Reflecting on our SLR results (cf. Section 4), guideline (cf. Section 5), and recommendations (cf. Section 6), we considered the most interesting question on financial incentives for researchers in SE experimentation to be: *How to design payoff functions for an SE experiment?* In the following, we discuss this question based on an example setup for an experiment on code reviews in which the participants have to find bugs in a piece of code. We extensively discussed this setup among the authors to understand payoff functions for SE. During these discussions, we essentially followed our guidelines and recommendations, resulting in an even more elaborate design of a bug-finding experiment aimed at comparing the impact of different financial incentivisation strategies—which has been peer reviewed and accepted as a registered report (Krüger et al. 2022). Our preliminary results already indicate that different payoff functions, and thus varying forms of financial incentives, impact the results of our experiment (which is otherwise identical). For simplicity, we present three short examples in which we incrementally increase the complexity of incentivizing.

Example 1: Bug Bounties Bounties for fixing bugs in open-source systems are a means to incentivize developers in practice (Krishnamurthy and Tripathi 2006). We now apply our flowchart in Fig. 2 to this example, analyzing whether applying a payoff function would be useful to improve the validity of an experiment. First, we acknowledge that the performance of developers is important for finding bugs (Q_1). It is further reasonable that the performance depends on the effort a developer spends (Q_2), since taking more time or concentration to search for a bug increases the probability of finding it. Now, we assess whether money can influence the effort (Q_3). The question is whether a larger bounty will attract more developers or make one developer spend more effort on the bug, while the complexity of the bug remains constant—which we assume to be the case. Next, we understand that bug bounties in the real world involve rewards in the form of money for those who detect the bug (Q_4). As a consequence, a payoff function would in general be useful in the context of our example. However, we now have to understand whether there are risks of crowding out not existing in the real world (Q_5) and habituation within our experiment (Q_6). Crowding out may theoretically happen, yet it likewise can occur in the real-world context of bug bounties. Habituation is not very likely, unless the very same experiment is repeated several times with the same participants. Therefore, we can conclude that it would be useful and recommended to use a payoff function for this example—assuming there is enough funding (Q_7) and paying participants is legal (Q_8).

This leads to the question what an appropriate payoff function for bug bounties could look like following our recommendations. While iterating through our guideline, we could already capture most of the real-world properties relevant for designing a payoff function (R_1), such as the tournament mode inherent to such bounties. Still, it has also become clear that it is almost impossible to conduct a laboratory experiment including the entire complexity of open-source systems, for instance, to capture all intrinsic motivators (R_7) and development methodologies (R_8). Thus, it is important to narrow down the precise problem and relevant parameters of the environment (R_{11}). The most basic abstraction could be: n individuals (i.e., developers) try to find a bug in a piece of code, and only the participant who finds it first will be rewarded. Simplified, to design a payoff function, all we need to know is the average time participants need to find a bug of chosen complexity, which can be obtained from pilot experiments, published empirical studies, or by analyzing version-control data (R_2). Knowing this expected time and the hourly wage of the participants, we can estimate the size of the bounty (R_6).

Concretely, consider students with an hourly wage for student assistants of \$12 per hour and an observed, average bug-fixing time of 30 minutes. For a between-subject experiment with three groups, we intend to invite 10 students. We aim to compensate each student with \$12 per hour on average (R_6), while ensuring the real-world tournament mode (R_1). So, we decide that all students receive a show-up fee (e.g., \$5). The first student in each group to find the bug receives the additional bounty of \$12 (i.e., winners-take-all tournament, cf. Table 1). If no student in a group finds the bug within the time of the experiment or all participants give up, nobody receives the bounty. This design fulfills all the criteria for a payoff function (cf. Section 2), mimics the real-world, and should imitate the participants' motivation to solve the task, and thus receive the additional reward of \$12. Lastly, we estimate (e.g., via a test run) that the bug will only be found in two of three groups, and estimate the expected average payoff per student, which corresponds to an hourly wage of \$11.60 (i.e., total amount paid to all participants: \$3 groups * 10 participants * \$5 + 2 * \$12 bounty = \$174, divided by the total number of students: $\$174/30 = 5.8$ for half an hour). So, the average payoff per student is very close to their real-world wage to improve their participation and is performance-dependent to imitate real-world effort. Moreover, based on the envisioned population size, we assess whether the overall costs are reasonable to achieve these goals (R_9). When documenting the experiment, we report all design decisions and the payoff function to allow others to replicate the experiment (R_{10}).

Like this, the general incentives of the bounty program are replicated in the lab. Being aware that this example abstracts from reality, it still enables researchers to incrementally include specific financial incentives of interest (R_3). For instance, it is possible to include a series of bug bounties (i.e., bugs of different complexity and bounty sizes) using the same method. So, it becomes possible to analyze selection processes, for example, to understand which participants decide to go for which bounty. Alternatively, we could allow for team formations. A group of developers is more likely to find a bug, yet would need to share the bounty (i.e., resembling a proportional-prize contest). All of this could be captured within a general framework of the experiment that allows for easy replication and extensions in different settings.

Example 2: Piece Rates As a second example, consider that an individual participant needs to identify as many *bugs* as possible after having received a report about failed tests. Every bug they find improves the code quality, which we consider to be another (Q_{4a}) non-monetary (Q_{4b}) incentive we can induce with financial incentives (Q_{4c}). Thus, it is reasonable to include a piece rate, for instance, per correctly identified bug a participant would receive a certain amount of money m—which again could be elicited in pilot studies or by reflecting on the negative impact in terms of the costs bugs cause. A payoff function could simply be: $\#bugs * m$.

Let us increase the complexity of this example in two more steps. First, the participant identifies something as a bug that is actually not a bug ($\#!bugs$). In the real world (R_1), this could induce costs (e.g., requiring additional code reviews). Thus, the experimenter may want to implement a penalty p in the payoff function to resemble the potential costs that may be caused in reality. In this case, the payoff function becomes: $\#bugs * m - \#!bugs * p$. Second, having missed a bug ($\#?bugs$) in the code is also costly (c), for instance, in terms of reduced customer satisfaction or damages caused by misbehavior of the software. Even if some of such consequences are carried by the company, the incentive environment for the individual developer is similar. For example, a developer indicating all lines of code as bugs or missing important bugs may decrease their reputation and chances for promotion or be required to work overtime. This can be implemented as: $\#bugs * m - \#!bugs * p - \#?bugs * c$. Now, it is up to the experimenter to properly tune the parameters m, p, and c in the right proportion

to each other (R_2), so that they mirror consequences for developers in the real world (e.g., $c > p$). In parallel, the parameters must still cover the opportunity costs of the participants. Given the financial incentives induced through the payoff function, it now becomes possible to analyze the effect of different aspects of interest, such as time pressure or cost-benefit ratios.

Example 3: Incentives as (In-)Dependent Variables On a final note, we remark that payoffs are clearly aligned with (at least one) dependent variable, and thus could themselves represent a dependent variable—which can occasionally be useful. Moreover, financial incentives can be a valuable addition to SE experiments when using them as independent variable. For instance, if developers receive payoffs for finding bugs correctly, but get penalized for the time they use (R_5), the payoff can serve as an independent variable to identify whether and how participants optimize for their payoff. Some participants may leave the experiment fast or just mark some bugs to maximize their payoff due to the missing penalty, while others are inclined to receive the bonus (R_4). Again, this issue requires researchers to scope the payoff function properly (e.g., avoiding that leaving immediately yields a higher payoff than finding bugs).

8 Threats to Validity

Next, we describe threats to the validity of our research. To this end, we follow the guidelines of Kitchenham et al. (2015) for reporting threats to SLRs.

Construct Validity The construct validity is concerned with how well the design of the study is able to address the research question. Primarily, we employed an SLR to capture the current state-of-the-art of using incentives in SE experimentation. So, a threat to construct validity is that the general content of the studies in our SLR was not on financial incentives. However, our investigation indicates that there are no studies in SE that investigate the role of financial incentives in SE experiments (cf. Section 4.2). Even though this may threaten the construct validity, then the best solution to investigate the use of financial incentives in SE is by connecting it to other disciplines that make use of such incentives. To derive our guideline and recommendations, we performed an interdisciplinary analysis together with researchers from experimental economics and behavioral psychology (Section 3.6). Both disciplines use financial incentives to different extents, and we relied on guidelines established in the respective communities (Harrison and List 2004; Weimann and Brosig-Koch 2019; van Dijk et al. 2001; Erkal et al. 2018; Weber and Camerer 2006; Kirk 2013). While experimental economics and behavioral psychology study the pros and cons of financial incentives, particularly experimental economics points out their benefits. Covering perspectives from these two disciplines helps us provide guidelines for SE researchers on when to use and when not to use financial incentives in experiments with human participants. So, while our interdisciplinary analysis may have introduced biases (e.g., due to personal experiences and knowledge), it helped us to more easily capture and understand pros, cons, and forms of financial incentives.

Internal Validity The internal validity is concerned with the conduct of our methodology. A first threat to our work are the missing details on whether and how incentives have been used in SE experimentation. Consequently, researchers may employ financial incentives just as we discussed, they just do not report them. Arguably, this is highly unlikely considering that existing guidelines in SE also rarely mention the use of financial incentives. Moreover, the general picture in our dataset clearly hints at the insight that mostly completion fees are

used. Nonetheless, this internal threat remains, but our contributions are relevant considering particularly the state-of-the-art on reporting incentives.

Another threat to the internal validity is that we may have missed important publications or data in the publications. Since we considered only a subset of all published experiments in SE, we may have missed relevant publications that employ and report incentivisation in far more detail or various different forms. While this threat definitely remains, we argue that our sample is feasible to tackle our goal of capturing the state-of-the-art, since we covered 11 of the most prominent high-quality publication venues for empirical SE in 2020 and 2021—analyzing 105 publications out of 2,284. If publications at these venues do not report on using financial incentives or mention according guidelines, we would not expect publications at other venues to do so. Moreover, the overall picture (i.e., only four experiments employing advanced payoff functions) is clearly supporting our subjective perception, and established guidelines do also rarely mention financial incentives. We remark again that we did not find any hints on the COVID-19 pandemic impacting how incentives have been used (Section 3.2), which should have impacted publications from 2021 at most. As described (cf. Section 3), we performed multiple rounds of validation that did not reveal major errors that would threaten our findings (i.e., completely wrong data entries), but only small inconsistencies (e.g., wrong number of participants, larger payments). Also, we reiterated multiple times through most publications during our analysis, particularly to provide context for the authors from experimental economics and behavioral psychology. These means mitigate this threat, but cannot fully prevent it.

We performed our interdisciplinary analysis through repeated discussions during which we defined and refined our guideline, recommendations, and examples. Since this also involved our personal experiences, the involved subjectivity threatens the internal validity of our analysis. We aimed to mitigate this threat by systematically eliciting the state-of-the-art in SE and checking guidelines from all involved disciplines. Moreover, all authors are experienced experimenters in their disciplines (e.g., the third author is member of the *Magdeburg Experimental Laboratory of Economic Research*[10]), which mitigates this threat. Still, we cannot fully overcome this threat, but it definitely does not invalidate our findings or guidelines—which seem needed considering the discrepancy between SE experimentation and other disciplines regarding the use of financial incentives.

External Validity The external validity is concerned with the extent to which our results can be generalized. Consequently, a threat to our SLR arises from the venues we selected for our manual search. As we discussed before, our overview of incentives in SE may be incomplete, since publications at other venues potentially focus more on this issue. Again, we considered the most prominent venues on empirical SE, checked established guidelines, and aimed to identify related work through automated searches. Overall, the picture remained the same, namely that financial incentives are sparsely used (or reported) in SE experimentation. Nonetheless, our results may not be fully generalizable, but this does not impair our actual contributions; seeing that even high-quality publications rarely report on the use of incentives.

A threat to the external validity of our contributions is that our guidelines and recommendations may not be fully transferable to all types of SE experiments. We mitigated this threat by capturing the state-of-the-art in SE, consulting guidelines, and discussing the potential designs of incentives for different SE experiments. This helped us to put our knowledge on financial incentives into the context of SE. Moreover, we actively discussed when not to use financial incentives and included such points into our guideline and our recommendations,

[10] http://maxlab.ovgu.de/en/

which basically incorporates this threat into our contributions as an experimental property researchers have to consider.

Conclusion Validity The conclusion validity is concerned with the degree to which our findings are credible (i.e., supported by the results). We personally reviewed the state-of-the-art in SE experimentation and existing guidelines from the different communities to derive our contributions. Consequently, other researchers may derive different insights or recommendations from our data. We mitigated this threat as stated before. Additionally, we provide an open-access repository with our dataset to enable other researchers to replicate and evaluate our work.[2]

A last threat to our guideline and recommendations is that we have not yet conducted a full-fledged experiment using them. However, we have derived them based on decades of research in other disciplines and have tested all steps when designing an experiment on comparing financial incentives in SE, with the design being registered and approved by reviewers (Krüger et al. 2022). Financial incentives could potentially behave differently in SE than in other disciplines. However, this is highly unlikely considering existing research, and it is rather a task to identify the SE experiments in which financial incentives are useful and in which they are not. Particularly for this task, we have provided our guideline and recommendations, and we argue that such threats to our contributions are minimal.

9 Conclusion

In this article, we discussed the use of financial incentives in SE experimentation. We captured to what extent incentives are used in SE based on an SLR, and involved researchers from two other disciplines to discuss whether and how financial incentives may be used in different settings of SE experiments. Our key contributions are:

- We found that financial incentives are rarely used to their full potential in SE experimentation, even though they could improve participation, motivation, validity, and analysis methods (Section 4)
- We defined a guideline comprising 11 questions that SE researchers should answer to decide whether to use advanced financial incentives (i.e., payoff functions) in their experiments or not (Section 5).
- We discussed 11 recommendations that can help SE researchers to design payoff functions according to their experimental setup (Section 6).
- We exemplified how to use our guideline and recommendations to help other researchers employ financial incentives in their experiments (Section 7).

Overall, we hope that our contributions help SE researchers understand, decide on, and use advanced financial incentives in their experiments, and that they will be incorporated into general purpose guidelines (e.g., ACM SIGSOFT Empirical Standards). In addition, we discussed various open questions and challenges that demand for further research.

Regarding future work, we plan to conduct experiments in which we compare the impact of different incentivisation strategies (Krüger et al. 2022). In particular, we aim to capture real-world settings in SE practice and aim to translate these into payoff functions, which we then can employ in experiments. These studies will help us extend our guidelines and provide additional recommendations for researchers on using incentives.

A – Overview of the Included Publications

In Table 3, we provide an overview of the publications we included in our literature review.

Table 3 The 105 publications we analyzed (top: 56 from 2020; bottom: 49 from 2021)

ref	method	design	participants				incentives		
			profile	#	valid	countries (ISO 3166)	form	value	costs
Abdellatif et al. (2020)	exp	w	s	12		CAN	—	—	—
Amálio et al. (2020)	exp	w	s	43		PRT, LUX, GBR	completion fee	50 € vouchers	2,150 €
Azizi et al. (2020)	exp	w	s	12		IRN	—	—	—
Bai et al. (2020)	os	n/a	s	24	(18)	USA	completion lottery	one $30 Amazon gift card	$30
Bao et al. (2020)	exp	b	s	10		CHN	—	—	—
Behroozi et al. (2020)	exp	b	s	50	(48)	USA	show-up fee	course credits	—
Beschastnikh et al. (2020)	exp	b	s, r	39		CAN, USA	—	—	—
Chattopadhyay et al. (2020)	os	n/a	d	10		USA	—	—	—
Cornejo et al. (2020)	exp	w	s, r	22		ITA	—	—	—
Corradini et al. (2020)	exp	b	s	26		ITA	—	—	—
Czepa and Zdun (2020)	exp	b	s	216		AUT	piece rate	course credits	—
Dias et al. (2020)	exp	w	d	16		CHL	—	—	—
Do et al. (2020)	exp	w	s, r	20		DEU	—	—	—
Fakhoury et al. (2020)	exp	hybrid	s	25		USA	show-up fee (?)	$15 gift cards	$375
Fucci et al. (2020)	exp	b	s	45		ITA	—	—	—
Gil et al. (2020)	os + exp	n/a + b	p + s	6 + 22		ESP	—	—	—
Girardi et al. (2020)	os	b	s	27		ITA	completion fee	meal voucher	—
Gopstein et al. (2020)	os	n/a	s, d	14		USA	—	—	—

Table 3 Continued

ref	method	design	participants			incentives		
			profile	#	valid	countries (ISO 3166)	form	value	costs
Goumopoulos and Mavrommati (2020)	os	n/a	r	20		GRC	—	—	—
Gralha et al. (2020)	exp	b	s, r, d	180		PRT	—	—	—
Huang et al. (2020)	exp	hybrid	s	37	(36)	USA	completion fee	$75 and 3D brain model	$2, 775
Jolak et al. (2020)	exp	b	s	240	(226)	SWE, DEU, FRA, SVK	—	—	—
Kafali et al. (2020)	exp	b	s	32		USA	completion fee	$20	$640
Krueger et al. (2020)	os	w	s	30	(24)	USA	completion fee	$75 and 3D brain model	$2, 250
LaToza et al. (2020)	exp	b	s, d	28		USA	show-up fee (?)	$30 gift cards	$840
Lian et al. (2020)	exp	b	d	10		CHN	—	—	—
Masood et al. (2020)	os	n/a	d	7		NZL	—	—	—
McChesney and Bond (2020)	os	b	d	28		GBR	—	—	—
Morales et al. (2020)	exp	w	d	30		CAN	—	—	—
Núñez et al. (2020)	os + exp	n/a + b	s, d	11 + 6		PRY	—	—	—
Peitek et al. (2020a)	exp	hybrid	d	31		DEU	—	—	—
Peitek et al. (2020b)	os	n/a	s	28		DEU	show-up fee (?)	20 €	560 €
Ren et al. (2020)	exp	w	s	54		ECU	—	—	—
Romano et al. (2020)	exp	b	s	83		ITA, USA	show-up fee (?)	$30, course credits	$510
Said et al. (2020)	exp	w	s, d	58		DEU	—	—	—
Satterfield et al. (2020)	os	n/a	s, d	17		CAN	—	—	—
Sayagh et al. (2020)	exp	b	s, d	55		CAN	completion fee (quality check)	35 CAD (5 of 7 d), course credits (s)	CAD 175

Table 3 Continued

ref	method	design	participants				incentives		
			profile	#	valid	countries (ISO 3166)	form	value	costs
Shargabi et al. (2020)	exp	w	s	178		MYS	winners-take-all tournament	—	—
Shen et al. (2020)	exp	b	n/a	8		CHN	—	—	—
Soltani et al. (2020)	exp	w	s	35		NLD	—	—	—
Spadini et al. (2020)	exp	b	d	243	(85)	CHE, NLD, SWE	completion fee	$5 donation to charity	$425
Stapleton et al. (2020)	exp	w	s, d	45		online	completion fee	$20	max. $900
Taipalus (2020)	exp	w	s	744		FIN	piece rate (?)	course credits	—
Tan and Li (2020)	exp	w	s	29	(27)	CHN	completion fee	course credits	—
Teixeira et al. (2020)	exp	b	d	30		BRA	—	—	—
Uddin et al. (2020)	exp	w	r, d	31		online, CAN, BGD	completion fee	$20 (d)	$360
Uesbeck et al. (2020)	exp	b	s, d	149	(109)	USA	completion fee	course credits	—
Urbieta et al. (2020)	exp	b	r, d	36		ARG, ESP	—	—	—
Valderas et al. (2020)	exp	w	r	9		ESP	—	—	—
Vassallo et al. (2020)	exp	w	s, d	17		CHE	—	—	—
Vegas et al. (2020)	exp	w	s	78	(62)	ESP	completion fee	course credits	—
Viticchié et al. (2020)	exp	b	s	87		ITA	show-up fee	course credits	—
Wang and Zhang (2020)	exp	b	s	280		USA	completion fee	$20	$5, 600
Yates et al. (2020)	os	n/a	r, d	27		CAN, US, GBR, Ireland	—	—	—
Zieris and Prechelt (2020)	os	n/a	d	14		DEU	—	—	—
Addazi and Ciccozzi (2021)	exp	w	r	18	(14)	SWE	—	—	—
Aghayi et al. (2021)	os	n/a	s, r	9		USA, GBR, ESP, IND	hourly wage	$20/hour (gift cards)	max. $630
Ahrens and Schneider (2021)	exp	b	s	29		DEU	show-up fee	course credits	—

Table 3 Continued

ref	method	design	participants				incentives		
			profile	#	valid	countries (ISO 3166)	form	value	costs
Alanazi et al. (2021)	os	n/a	d	18		online	—	—	—
Alhamed and Storer (2021)	exp	n/a	d	807		online	completion fee	money	—
Ampatzoglou et al (2021)	os	n/a	d	4		SWE	—	—	—
Baldassarre et al. (2021)	exp	w	s	69		ITA	completion fee	course credits	—
Blanco and Lucrédio (2021)	os	n/a	d, s	9		BRA	—	—	—
Braz et al. (2021)	exp	w	s, r, d	194	(146)	online	completion fee (quality check)	$5 donation to charity	$730
Cates et al. (2021)	exp	hybrid	d	191	(113)	online	—	—	—
Caulo et al. (2021)	os	n/a	s	31		ITA	—	—	—
da Costa et al. (2021)	exp	w	s, r	64		BRA	—	—	—
Dalpiaz et al. (2021)	exp	hybrid	s	150	(142)	ISR, NLD	completion fee	course credits	—
Danilova et al. (2021)	os	n/a	s, d, p	249		online, DEU, AUT, GBR, USA, ESP, ITA	show-up (s, p) / completion (s, d) fee	10 (d)/5 (s)/1 course credits (s)	490 € (p), 250 € (s), 125 € (d)
Echeverría et al. (2021)	exp	b	d	10		ESP	—	—	—
Endres et al. (2021a)	exp	b	s	97	(57)	USA	completion fee	$20 per session (max 11)	max. $12, 540
Endres et al. (2021b)	exp	w	s	37	(31, 23)	USA	completion fee	2*$20	$1, 200
Foundjem et al. (2021)	os	n/a	d	72		DEU	—	—	—
Guerriero et al. (2021)	exp	w	p	23		Europe	—	—	—
Hallett et al. (2021)	exp	b	d	138		online	completion fee	£5	£690
Jørgensen et al. (2021)	os	n/a	d	104		POL, UKR	hourly wages	money	—

Table 3 Continued

ref	method	design	participants				incentives		
			profile	#	valid	countries (ISO 3166)	form	value	costs
Jørgensen and Grov (2021)	exp	hybrid	d	52		NOR	hourly wages, contracts	$180—4,000	—
Karac et al. ((2021))	exp	w	s	52	(48)	FIN	—	—	—
Kifetew et al. (2021)	exp	hybrid	s	40		DEU, ITA	—	—	—
Kirby et al. (2021)	os	n/a	d	10		CAN, HRV, DEU, NOR, AUS, USA	—	—	—
Kuttal et al. (2021)	exp	w	r, d, p	10		n/a	—	—	—
Lavalle et al. (2021)	exp	w	s, d	97		ESP	completion fee (s)	course credits (s)	—
Liu et al. (2021)	exp	hybrid	s	10		CHN	contracts	money	—
Melo et al. (2021)	os	n/a	s, d	>10		BRA	—	—	—
Meyer et al. (2021)	os	n/a	d	59	(52)	USA, CAN, CHE, BRA	completion fee	$50 Amazon git card	max. $2, 950
Mohanani et al. (2021)	exp	b	s	76		GBR, FIN	completion fee	meal voucher, course credits	—
Muntean et al. (2021)	exp	b	s	30		DEU, SWE, CHN	—	—	—
Olsson et al. (2021)	exp	b	d	40		SWE	—	—	—
Ore et al. (2021)	exp	w	d	97		online	completion fee	$2, $10	$1, 164
Panach et al. (2021)	exp	w	s	78		CHL, ESP	—	—	—
Paulweber et al. (2021a)	exp	b	s	105		AUT	piece rate	up to 6 course credits	—
Paulweber et al. (2021b)	exp	b	s	98		AUT	piece rate	up to 6 course credits	—
Peitek et al. (2021)	os	n/a	s	19	(18)	DEU	show-up fee (?)	money	—
Santos et al. (2021)	exp	hybrid	s, d	411		Europe	—	—	—

Table 3 Continued

ref	method	design	participants				incentives		
			profile	#	valid	countries (ISO 3166)	form	value	costs
Saputri and Lee (2021)	exp	hybrid	s, d	30		KOR	—	—	—
Scalabrino et al. (2021)	os	n/a	s, d	63		online	—	—	—
Scoccia et al. (2021)	os + exp	n/a + w	d + s, p	11 + 47		online	—	—	—
Sharafi et al. (2021)	exp	n/a	s	111	(99)	USA	completion fee	money, course credits	—
Shen et al. (2021)	exp	b	s, d	8		CHN	—	—	—
Taipalus et al. (2021)	exp	b	s	152		FIN	show-up fee	course credits	—
Tosun et al. (2021)	exp	w	d	18	(17)	likely EST	—	—	—
Uddin et al. (2021)	exp	b	s, d	31		online, CAN, BGD	completion fee	$20 (d)	$360
Wiese et al. (2021)	os	n/a	s	143	(125)	USA	completion fee	$5, course credits	$120 (24 s)
Wyrich et al. (2021)	exp	b	s	45	(43)	DEU	—	—	—

Author Contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by Jacob Krüger and Gül Çalıklı. The data analysis, writing, and revising was done by all authors in collaboration. All authors read and approved the final manuscript.

Funding We received partial financial support for this research by the Otto-von-Guericke University Magdeburg Innovation Fund.

Data Availability The datasets generated and/or analyzed during this study will be made available in a Zenodo repository.[2]

Declarations

Competing Interests The authors have neither financial nor non-financial interests to declare.

Ethics Approval Does not apply.

References

Abdellatif A, Badran K, Shihab E (2020) MSRBot: Using bots to answer questions from software repositories. Empirical Software Engineering 25(3)

Addazi L, Ciccozzi F (2021) Blended graphical and textual modelling for UML profiles: a proof-of-concept implementation and experiment. J Syst Softw 175:110912. https://doi.org/10.1016/j.jss.2021.110912

Aghayi E, LaToza TD, Surendra P, Abolghasemi S (2021) Crowdsourced behavior-driven development. J Syst Softw 171:110840. https://doi.org/10.1016/j.jss.2020.110840

Aguinis H, Villamor I, Ramani RS (2021) MTurk research: review and recommendations. Journal of Management, 47(4)

Ahrens M, Schneider K (2021) Improving requirements specification use by transferring attention with eye tracking data. Inf Softw Technol 131:106483. https://doi.org/10.1016/j.infsof.2020.106483

Alanazi R, Gharibi G, Lee Y (2021) Facilitating program comprehension with call graph multilevel hierarchical abstractions. J Syst Softw 176:110945. https://doi.org/10.1016/j.jss.2021.110945

Alhamed M, Storer T (2021) Playing planning poker in crowds: human computation of software effort estimates. In: International Conference on Software Engineering (ICSE), IEEE, pp 1–12. https://doi.org/10.1109/ICSE43902.2021.00014

Allodi L, Cremonini M, Massacci F, Shim W (2020) Measuring the accuracy of software vulnerability assessments: experiments with students and professionals. Empirical Software Engineering 25(2)

Amálio N, Briand LC, Kelsen P (2020) An experimental scrutiny of visual design modelling: VCL up against UML+OCL. Empirical Software Engineering 25(2)

Ampatzoglou A, Arvanitou E, Ampatzoglou A, Avgeriou P, Tsintzira A, Chatzigeorgiou A (2021) Architectural decision-making as a financial investment: an industrial case study. Inf Softw Technol 129:106412. https://doi.org/10.1016/j.infsof.2020.106412

Amrhein V, Greenland S, McShane B (2019) Retire statistical significance: scientists rise up against statistical significance. Nature 567(7748):305–307. https://doi.org/10.1038/d41586-019-00857-9

Azizi B, Zamani B, Rahimi SK (2020) SEET: symbolic execution of ETL transformations. J Syst Softw 168. https://doi.org/10.1016/j.jss.2020.110675

Bai GR, Kayani J, Stolee KT (2020) How graduate computing students search when using an unfamiliar programming language. In: International Conference on Program Comprehension (ICPC), ACM, pp 160–171. https://doi.org/10.1145/3387904.3389274

Baker M (2016) Statisticians issue warning over misuse of p values. Nature 531(7593):151–151. https://doi.org/10.1038/nature.2016.19503
Baldassarre MT, Caivano D, Fucci D, Juristo N, Romano S, Scanniello G, Turhan B (2021) Studying test-driven development and its retainment over a six-month time span. J Syst Softw 176:110937. https://doi.org/10.1016/j.jss.2021.110937
Baltussen G, Post GT, van den Assem MJ, Wakker PP (2012) Random incentive systems in a dynamic choice experiment. Exp Econ 15(3):418–443
Bao L, Xing Z, Xia X, Lo D, Wu M, Yang X (2020) psc2code: denoising code extraction from programming screencasts. ACM Transactions on Software Engineering and Methodology, 29(3):21:1–21:38. https://doi.org/10.1145/3392093
Becker R, Möser S, Glauser D (2019) Cash vs. vouchers vs. gifts in web surveys of a mature panel study-main effects in a long-term incentives experiment across three panel waves. Soc Sci Res 81:221–234
Behroozi M, Shirolkar S, Barik T, Parnin C (2020) Does stress impact technical interview performance? In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), ACM, pp 481–492. https://doi.org/10.1145/3368089.3409712
Beschastnikh I, Liu P, Xing A, Wang P, Brun Y, Ernst MD (2020) Visualizing distributed system executions. ACM Transactions on Software Engineering and Methodology, 29(2):9:1–9:38. https://doi.org/10.1145/3375633
Blanco JZ, Lucrédio D (2021) A holistic approach for cross-platform software development. J Syst Softw 179:110985. https://doi.org/10.1016/j.jss.2021.110985
Braz L, Fregnan E, Çalikli G, Bacchelli A (2021) Why don't developers detect improper input validation? '; DROP TABLE Papers; –. In: International Conference on Software Engineering (ICSE), IEEE, pp 499–511. https://doi.org/10.1109/ICSE43902.2021.00054
Brüggen E, Wetzels M, De Ruyter K, Schillewaert N (2011) Individual differences in motivation to participate in online panels: the effect on reponse rate and reponse quality perceptions. Int J Mark Res 53(3):369–390
Bull C, Schotter A, Weigelt K (1987) Tournaments and piece rates: an experimental study. J Polit Econ 95(1):1–33
Burtch G, Hong Y, Bapna R, Griskevicius V (2018) Stimulating online reviews by combining financial incentives and social norms. Management Science, 64(5)
Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: a review and capital-labor-production framework. J Risk Uncertain 19(1):7–42
Camerer CF, Mobbs D (2017) Differences in behavior and brain activity during hypothetical and real choices. Trends Cogn Sci 21(1):46–56
Carpenter J, Huet-Vaughn E (2019) Real-effort tasks. Handbook of Research Methods and Applications in Experimental Economics
Carver JC, Jaccheri L, Morasca S, Shull F (2010) A checklist for integrating student empirical studies with research and teaching goals. Empir Softw Eng 15:35–59
Cason TN, Masters WA, Sheremeta RM (2010) Entry Into Winner-Take-All and Proportional-Prize Contests: An Experimental Study. J Public Econ 94(9–10):604–611
Cates R, Yunik N, Feitelson DG (2021) Does code structure affect comprehension? On using and naming intermediate variables. In: International Conference on Program Comprehension (ICPC), IEEE, pp 118–126. https://doi.org/10.1109/ICPC52881.2021.00020
Caulo M, Francese R, Scanniello G, Tortora G (2021) Relationships between personality traits and productivity in a multi-platform development context. In: International Conference on Evaluation and Assessment in Software Engineering (EASE), ACM, pp 70–79. https://doi.org/10.1145/3463274.3463327
Cerasoli CP, Nicklin JM, Ford MT (2014) Intrinsic motivation and extrinsic incentives jointly predict performance: a 40-year meta-analysis. Psychol Bull 140(4):980–1008. https://doi.org/10.1037/a0035661
Chattopadhyay S, Nelson N, Au A, Morales N, Sanchez C, Pandita R, Sarma A (2020) A tale from the trenches: cognitive biases and software development. In: International Conference on Software Engineering (ICSE), ACM, pp 654–665. https://doi.org/10.1145/3377811.3380330
Cornejo O, Briola D, Micucci D, Mariani L (2020) In-the-field monitoring of functional calls: is it feasible? J Syst Softw 163. https://doi.org/10.1016/j.jss.2020.110523
Corradini F, Morichetta A, Polini A, Re B, Rossi L, Tiezzi F (2020) Correctness checking for BPMN collaborations with sub-processes. J Syst Softw 166. https://doi.org/10.1016/j.jss.2020.110594
da Costa JAS, Gheyi R, Ribeiro M, Apel S, Alves V, Fonseca B, Medeiros F, Garcia A (2021) Evaluating refactorings for disciplining #Ifdef annotations: an eye tracking study with novices. Empir Softw Eng 26(5):92. https://doi.org/10.1007/s10664-021-10002-8
Cubitt RP, Starmer C, Sugden R (1998) On the validity of the random lottery incentive system. Exp Econ 1(2):115–131

Czepa C, Zdun U (2020) On the understandability of temporal properties formalized in linear temporal logic, property specification patterns and event processing language. IEEE Trans Software Eng 46(1):100–112. https://doi.org/10.1109/TSE.2018.2859926

Dalpiaz F, Gieske P, Sturm A (2021) On deriving conceptual models drom user requirements: an empirical study. Inf Softw Technol 131:106484. https://doi.org/10.1016/j.infsof.2020.106484

Danilova A, Naiakshina A, Horstmann S, Smith M (2021) Do you really code? designing and evaluating screening questions for online surveys with programmers. In: International Conference on Software Engineering (ICSE), IEEE, pp 537–548. https://doi.org/10.1109/ICSE43902.2021.00057

David MC, Ware RS (2014) Meta-analysis of randomized controlled trials supports the use of incentives for inducing response to electronic health surveys. J Clin Epidemiol 67(11):1210–1221

Deci EL (1971) Effects of externally mediated rewards on intrinsic motivation. J Pers Soc Psychol 18(1):105

DellaVigna S, Pope D (2018) What motivates effort? Evidence and expert forecasts. Rev Econ Stud 85(2):1029–1069

Dias M, Orellana D, Vidal SA, Merino L, Bergel A (2020) Evaluating a Visual Approach for Understanding JavaScript Source Code. In: International Conference on Program Comprehension (ICPC), ACM, pp 128–138. https://doi.org/10.1145/3387904.3389275

van Dijk F, Sonnemans J, van Winden F (2001) Incentive systems in a real effort experiment. Eur Econ Rev 45(2):187–214. https://doi.org/10.1016/s0014-2921(00)00056-8

Do LNQ, Krüger S, Hill P, Ali K, Bodden E (2020) Debugging static analysis. IEEE Trans Software Eng 46(7):697–709. https://doi.org/10.1109/TSE.2018.2868349

Echeverría J, Pérez F, Panach JI, Cetina C (2021) An empirical study of performance using clone & own and software product lines in an industrial context. Inf Softw Technol 130:106444. https://doi.org/10.1016/j.infsof.2020.106444

Edwards P, Cooper R, Roberts I, Frost C (2005) Meta-analysis of randomised trials of monetary incentives and response to mailed questionnaires. Journal of Epidemiology & Community Health 59(11):987–999

Endres M, Fansher M, Shah P, Weimer W (2021a) To read or to rotate? Comparing the effects of technical reading training and spatial skills training on novice programming ability. In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), ACM, pp 754–766. https://doi.org/10.1145/3468264.3468583

Endres M, Karas Z, Hu X, Kovelman I, Weimer W (2021b) Relating reading, visualization, and coding for new programmers: a neuroimaging study. In: International Conference on Software Engineering (ICSE), IEEE, pp 600–612. https://doi.org/10.1109/ICSE43902.2021.00062

Erkal N, Gangadharan L, Koh BH (2018) Monetary and non-monetary incentives in real-effort tournaments. Eur Econ Rev 101:528–545. https://doi.org/10.1016/j.euroecorev.2017.10.021

Esteves-Sorenson C, Broce R (2020) Do monetary incentives undermine performance on intrinsically enjoyable tasks? A field test. Review of Economics and Statistics pp 1–46

Fakhoury S, Roy D, Ma Y, Arnaoudova V, Adesope OO (2020) Measuring the impact of lexical and structural inconsistencies on developers' cognitive load during bug localization. Empir Softw Eng 25(3):2140–2178. https://doi.org/10.1007/s10664-019-09751-4

Felderer M, Travassos GH (2020) Contemporary empirical methods in software engineering. Springer. https://doi.org/10.1007/978-3-030-32489-6

Feltovich N (2011) What's to know about laboratory experimentation in economics? Journal of Economic Surveys 25(2):371–379

Fiore AT, Cheshire C, Taylor L, Mendelsohn GA (2014) Incentives to participate in online research: an experimental examination of "surprise" incentives. In: International Conference on Human Factors in Computing Systems (CHI), ACM, pp 3433–3442. https://doi.org/10.1145/2556288.2557418

Foundjem A, Eghan EE, Adams B (2021) Onboarding vs. diversity, productivity, and quality - empirical study of the Openstack ecosystem. In: International Conference on Software Engineering (ICSE), IEEE, pp 1033–1045. https://doi.org/10.1109/ICSE43902.2021.00097

Frey BS (1997) Not Just for the Money. Edward Elgar Publishing

Fucci D, Scanniello G, Romano S, Juristo N (2020) Need for sleep: the impact of a night of sleep deprivation on novice developers' performance. IEEE Trans Software Eng 46(1):1–19. https://doi.org/10.1109/TSE.2018.2834900

Gil M, Albert M, Fons J, Pelechano V (2020) Engineering human-in-the-loop interactions in cyber-physical systems. Inf Softw Technol 126. https://doi.org/10.1016/j.infsof.2020.106349

Girardi D, Novielli N, Fucci D, Lanubile F (2020) Recognizing developers' emotions while programming. In: International Conference on Software Engineering (ICSE), ACM, pp 666–677. https://doi.org/10.1145/3377811.3380374

Glasgow MJ, Murphy MS (1992) An experiment with financial incentives for a small software development team. In: Washington Ada Symposium on Ada: Empowering Software Users and Developers (WADAS), ACM, pp 86–92. https://doi.org/10.1145/257683.257713
Gneezy U, Rustichini A (2000) Pay enough or don't pay at all. Q J Econ 115(3):791–810
Gopstein D, Fayard A, Apel S, Cappos J (2020) Thinking aloud about confusing code: a qualitative investigation of program comprehension and atoms of confusion. In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), ACM, pp 605–616. https://doi.org/10.1145/3368089.3409714
Goumopoulos C, Mavrommati I (2020) A framework for pervasive computing applications based on smart objects and end user development. J Syst Softw 162. https://doi.org/10.1016/j.jss.2019.110496
Gralha C, Goulão M, Araújo J (2020) Are there gender differences when interacting with social goal models? Empir Softw Eng 25(6):5416–5453. https://doi.org/10.1007/s10664-020-09883-y
Grossklags J (2007) Experimental economics and experimental computer science: a survey. In: Workshop on Experimental Computer Science (ExpCS), ACM, https://doi.org/10.1145/1281700.1281711
Guerriero M, Tamburri DA, Nitto ED (2021) StreamGen: model-driven development of distributed streaming applications. ACM Transactions on Software Engineering and Methodology 30(1):1:1–1:30. https://doi.org/10.1145/3408895
Gunasti K, Baskin E (2018) Is a $200 nordstrom gift card worth more or less than a $200 gap gift card? the asymmetric valuations of luxury gift cards. J Retail 94(4):380–392
Hallett J, Patnaik N, Shreeve B, Rashid A (2021) "Do this! do that!, and nothing will happen" do specifications lead to securely stored passwords? In: IEEE, IEEE, pp 486–498. https://doi.org/10.1109/ICSE43902.2021.00053
Harrison GW (1992) Theory and misbehavior of first-price auctions: Reply. Am Econ Rev 82(5):1426–1443
Harrison GW, List JA (2004) Field experiments. Journal of Economic literature 42(4):1009–1055
Harrison GW, Lau MI, Rutström EE (2009) Risk attitudes, randomization to treatment, and self-selection into experiments. Journal of Economic Behavior & Organization 70(3):498–507
Hertwig R, Ortmann A (2001) Experimental practices in economics: a methodological challenge for psychologists? Behavioral and Brain Sciences 24(3):383–403
Ho CJ, Slivkins A, Suri S, Vaughan JW (2015) Incentivizing high quality crowdwork. In: International Conference on World Wide Web (WWW), WWW Conference, pp 419–429. https://doi.org/10.1145/2736277.2741102
Höst M, Wohlin C, Thelin T (2005) Experimental context classification. In: International Conference on Software Engineering (ICSE), ACM, pp 470–478. https://doi.org/10.1145/1062455.1062539
Huang Y, Leach K, Sharafi Z, McKay N, Santander T, Weimer W (2020) Biases and differences in code review using medical imaging and eye-tracking: genders, humans, and machines. In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), ACM, pp 456–468. https://doi.org/10.1145/3368089.3409681
Jolak R, Savary-Leblanc M, Dalibor M, Wortmann A, Hebig R, Vincur J, Polásek I, Pallec XL, Gérard S, Chaudron MRV (2020) Software engineering whispers: the effect of textual vs. graphical software design descriptions on software design communication. Empirical Software Engineering 25(6):4427–4471. https://doi.org/10.1007/s10664-020-09835-6
Jørgensen M, Grov J (2021) A field experiment on trialsourcing and the effect of contract types on outsourced software development. Inf Softw Technol 134:106559. https://doi.org/10.1016/j.infsof.2021.106559
Jørgensen M, Bergersen GR, Liestøl K (2021) Relations between effort estimates, skill indicators, and measured programming skill. IEEE Trans Software Eng 47(12):2892–2906. https://doi.org/10.1109/TSE.2020.2973638
Juristo N, Moreno AM (2001) Basics of software engineering experimentation. Springer. https://doi.org/10.1007/978-1-4757-3304-4
Kafali Ö, Ajmeri N, Singh MP (2020) DESEN: Specification of sociotechnical systems via patterns of regulation and control. ACM Transactions on Software Engineering and Methodology, 29(1):7:1–7:50. https://doi.org/10.1145/3365664
Kang MJ, Rangel A, Camus M, Camerer CF (2011) Hypothetical and real choice differentially activate common valuation areas. J Neurosci 31(2):461–468
Karac I, Turhan B, Juristo N (2021) A controlled experiment with novice developers on the impact of task description granularity on software quality in test-driven development. IEEE Trans Software Eng 47(7):1315–1330. https://doi.org/10.1109/TSE.2019.2920377
Karras O, Schneider K, Fricker SA (2020) Representing software project vision by means of video: a quality model for vision videos. J Syst Softw 162. https://doi.org/10.1016/j.jss.2019.110479
Kettles D, St Louis R, Steinbart P (2017) An experimental investigation of the individual and joint effects of financial and non-financial lnc entives on knowledge sharing using enterprise social media. Commu-

nications of the Association for Information Systems, 41(1):639–673. https://doi.org/10.17705/1CAIS.04127
Keusch F (2015) Why do people participate in web surveys? Applying survey participation theory to internet survey data collection. Management Review Quarterly 65(3):183–216
Kifetew FM, Perini A, Susi A, Siena A, Muñante D, Morales-Ramirez I (2021) Automating user-feedback driven requirements prioritization. Inf Softw Technol 138:106635. https://doi.org/10.1016/j.infsof.2021.106635
Kirby LJ, Boerstra E, Anderson ZJC, Rubin J (2021) Weighing the evidence: on relationship types in microservice extraction. In: International Conference on Program Comprehension (ICPC), IEEE, pp 358–368. https://doi.org/10.1109/ICPC52881.2021.00041
Kirk RE (2013) Experimental design: procedures for the behavioral sciences. Sage. https://doi.org/10.4135/9781483384733
Kitchenham BA, Budgen D, Brereton OP (2015) Evidence-Based Software Engineering and Systematic Reviews. CRC Press. https://doi.org/10.1201/b19467
Ko AJ, LaToza TD, Burnett MM (2015) A practical guide to controlled experiments of software engineering tools with human participants. Empir Softw Eng 20(1):110–141. https://doi.org/10.1007/s10664-013-9279-3
Krishnamurthy S, Tripathi AK (2006) Bounty programs in free/libre/open source software. In: The Economics of Open Source Software Development, Elsevier, pp 165–183
Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol 5(3):213–236
Krueger R, Huang Y, Liu X, Santander T, Weimer W, Leach K (2020) Neurological divide: an fMRI study of prose and code writing. In: International Conference on Software Engineering (ICSE), ACM, pp 678–690. https://doi.org/10.1145/3377811.3380348
Krüger J, Lausberger C, von Nostitz-Wallwitz I, Saake G, Leich T (2020) Search. review. repeat? An empirical study of threats to replicating SLR searches. Empirical Software Engineering, 25(1):627–677. https://doi.org/10.1007/s10664-019-09763-0
Krüger J, Çalıklı G, Bershadskyy D, Heyer R, Zabel S, Otto S (2022) Registered report: a laboratory experiment on using different financial-incentivization Sshemes in software-engineering experimentation. CoRR pp 1–10. https://doi.org/10.48550/arXiv.2202.10985
Kuttal SK, Chen X, Wang Z, Balali S, Sarma A (2021) Visual resume: exploring developers' online contributions for hiring. Inf Softw Technol 138:106633. https://doi.org/10.1016/j.infsof.2021.106633
LaToza TD, Arab M, Loksa D, Ko AJ (2020) Explicit programming strategies. Empir Softw Eng 25(4):2416–2449. https://doi.org/10.1007/s10664-020-09810-1
Lavalle A, Maté A, Trujillo J, Teruel MA, Rizzi S (2021) A methodology to automatically translate user requirements into visualizations: experimental validation. Inf Softw Technol 136:106592. https://doi.org/10.1016/j.infsof.2021.106592
Lian X, Liu W, Zhang L (2020) Assisting engineers extracting requirements on components from domain documents. Inf Softw Technol 118. https://doi.org/10.1016/j.infsof.2019.106196
Liu S, Li H, Jiang Z, Li X, Liu F, Zhong Y (2021) Rigorous code review by reverse engineering. Inf Softw Technol 133:106503. https://doi.org/10.1016/j.infsof.2020.106503
Locke EA, Schattke K (2019) Intrinsic and extrinsic motivation: time for expansion and clarification. Motivation Science 5(4):277–290. https://doi.org/10.1037/mot0000116
Marcus B, Schütz A (2005) Who are the people reluctant to participate in research? Personality correlates of four different types of nonresponse as inferred from self-and observer ratings. J Pers 73(4):959–984
Mason W, Watts DJ (2009) Financial incentives and the "performance of crowds". In: Workshop on Human Computation (HCOMP), ACM, pp 77–85
Masood Z, Hoda R, Blincoe K (2020) How agile teams make self-assignment work: a grounded theory study. Empir Softw Eng 25(6):4962–5005. https://doi.org/10.1007/s10664-020-09876-x
McChesney IR, Bond RR (2020) Observations on the linear order of program code reading patterns in programmers with dyslexia. In: International Conference on Evaluation and Assessment in Software Engineering (EASE), ACM, pp 81–89. https://doi.org/10.1145/3383219.3383228
Melo L, Wiese I, d'Amorim M (2021) Using docker to assist Q&A forum users. IEEE Trans Software Eng 47(11):2563–2574. https://doi.org/10.1109/TSE.2019.2956919
Merlo A, Schotter A (1992) Theory and misbehavior of first-price auctions: comment. Am Econ Rev 82(5):1413–1425
Meyer AN, Murphy GC, Zimmermann T, Fritz T (2021) Enabling good work habits in software developers through reflective goal-setting. IEEE Trans Software Eng 47(9):1872–1885. https://doi.org/10.1109/TSE.2019.2938525

Mohanani R, Turhan B, Ralph P (2021) Requirements framing affects design creativity. IEEE Trans Software Eng 47(5):936–947. https://doi.org/10.1109/TSE.2019.2909033
Moldovanu B, Sela A (2001) The optimal allocation of prizes in contests. Am Econ Rev 91(3):542–558
Morales R, Khomh F, Antoniol G (2020) RePOR: mimicking humans on refactoring tasks. Are we there yet? Empirical Software Engineering 25(4):2960–2996. https://doi.org/10.1007/s10664-020-09826-7
Muntean P, Monperrus M, Sun H, Grossklags J, Eckert C (2021) IntRepair: informed repairing of integer overflows. IEEE Trans Software Eng 47(10):2225–2241. https://doi.org/10.1109/TSE.2019.2946148
Murayama K, Matsumoto M, Izuma K, Matsumoto K (2010) Neural basis of the undermining effect of monetary reward on intrinsic motivation. Proc Natl Acad Sci 107(49):20911–20916
Nafi KW, Roy B, Roy CK, Schneider KA (2020) A universal cross language software similarity detector for open source software categorization. J Syst Softw 162. https://doi.org/10.1016/j.jss.2019.110491
Núñez M, Bonhaure D, González M, Cernuzzi L (2020) A model-driven approach for the development of native mobile applications focusing on the data layer. J Syst Softw 161. https://doi.org/10.1016/j.jss.2019.110489
Olsson J, Risfelt E, Besker T, Martini A, Torkar R (2021) Measuring affective states from technical debt. Empir Softw Eng 26(5):105. https://doi.org/10.1007/s10664-021-09998-w
Ore J, Detweiler C, Elbaum SG (2021) An empirical study on type annotations: accuracy, speed, and suggestion effectiveness. ACM Transactions on Software Engineering and Methodology 30(2):20:1–20:29. https://doi.org/10.1145/3439775
Paltenghi M, Pradel M (2021) Thinking like a developer? comparing the attention of humans with neural models of code. In: International Conference on Automated Software Engineering (ASE), IEEE, pp 867–879. https://doi.org/10.1109/ASE51524.2021.9678712
Panach JI, Dieste O, Marín B, España S, Vegas S, Pastor O, Juristo N (2021) Evaluating model-driven development claims with respect to quality: a family of experiments. IEEE Trans Software Eng 47(1):130–145. https://doi.org/10.1109/TSE.2018.2884706
Parco JE, Rapoport A, Stein WE (2002) Effects of financial incentives on the breakdown of mutual trust. Psychol Sci 13(3):292–297
Paulweber P, Simhandl G, Zdun U (2021) On the understandability of language constructs to structure the state and behavior in abstract state machine specifications: a controlled experiment. J Syst Softw 178:110987. https://doi.org/10.1016/j.jss.2021.110987
Paulweber P, Simhandl G, Zdun U (2021b) Specifying with interface and trait abstractions in abstract state machines: a controlled experiment. ACM Transactions on Software Engineering and Methodology 30(4):47:1–47:29. https://doi.org/10.1145/3450968
Peitek N, Siegmund J, Apel S (2020a) What drives the reading order of programmers? an eye tracking study. In: International Conference on Program Comprehension (ICPC), ACM, pp 342–353. https://doi.org/10.1145/3387904.3389279
Peitek N, Siegmund J, Apel S, Kästner C, Parnin C, Bethmann A, Leich T, Saake G, Brechmann A (2020) A look into programmers' heads. IEEE Trans Software Eng 46(4):442–462. https://doi.org/10.1109/TSE.2018.2863303
Peitek N, Apel S, Parnin C, Brechmann A, Siegmund J (2021) Program comprehension and code complexity metrics: an fMRI study. In: International Conference on Software Engineering (ICSE), IEEE, pp 524–536. https://doi.org/10.1109/ICSE43902.2021.00056
Petersen K, Gencel C (2013) Worldviews, research methods, and their relationship to validity in empirical software engineering research. In: Joint Conference of the International Workshop on Software Measurement (IWSM) and the International Conference on Software Process and Product Measurement (Mensura), IEEE, pp 81–89. https://doi.org/10.1109/iwsm-mensura.2013.22
Petersen K, Wohlin C (2009) Context in industrial software engineering research. In: International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE, pp 401–404
Pforr K (2015) Incentives: GESIS Survey Guidelines. Tech. rep, Leibniz Institute for the Social Sciences
Ralph P (2021) ACM SIGSOFT empirical standards released. ACM SIGSOFT Software Engineering Notes 46(1)
Rao M, Bacon DF, Parkes DC, Seltzer MI (2020) Incentivizing deep fixes in software economies. IEEE Trans Software Eng 46(1):51–70. https://doi.org/10.1109/TSE.2018.2842188
Ren R, Castro JW, Santos A, Pérez-Soler S, Acuña ST, de Lara J (2020) Collaborative modelling: chatbots or on-line tools? an experimental study. In: international conference on evaluation and assessment in software engineering (EASE), ACM, pp 260–269. https://doi.org/10.1145/3383219.3383246
Romano S, Vendome C, Scanniello G, Poshyvanyk D (2020) A multi-study investigation into dead code. IEEE Trans Software Eng 46(1):71–99. https://doi.org/10.1109/TSE.2018.2842781
Rydval O, Ortmann A (2004) How financial incentives and cognitive abilities affect task performance in laboratory settings: an illustration. Econ Lett 85(3):315–320

Różyńska J (2022) The ethical anatomy of payment for research participants. Med Health Care Philos 25(3):449–464
Said W, Quante J, Koschke R (2020) Mining understandable state machine models from embedded code. Empir Softw Eng 25(6):4759–4804. https://doi.org/10.1007/s10664-020-09865-0
Santos A, Vegas S, Dieste O, Uyaguari F, Tosun A, Fucci D, Turhan B, Scanniello G, Romano S, Karac I, Kuhrmann M, Mandic V, Ramac R, Pfahl D, Engblom C, Kyykka J, Rungi K, Palomeque C, Spisak J, Oivo M, Juristo N (2021) A family of experiments on test-driven development. Empir Softw Eng 26(3):42. https://doi.org/10.1007/s10664-020-09895-8
Saputri TRD, Lee S (2021) Integrated framework for incorporating sustainability design in software engineering life-cycle: an empirical study. Inf Softw Technol 129:106407. https://doi.org/10.1016/j.infsof.2020.106407
Satterfield C, Fritz T, Murphy GC (2020) Identifying and describing information seeking tasks. In: International Conference on Automated Software Engineering (ASE), IEEE, pp 797–808. https://doi.org/10.1145/3324884.3416537
Sayagh M, Kerzazi N, Petrillo F, Bennani K, Adams B (2020) What should your run-time configuration framework do to help developers? Empir Softw Eng 25(2):1259–1293. https://doi.org/10.1007/s10664-019-09790-x
Scalabrino S, Bavota G, Vendome C, Linares-Vásquez M, Poshyvanyk D, Oliveto R (2021) Automatically assessing code understandability. IEEE Trans Software Eng 47(3):595–613. https://doi.org/10.1109/TSE.2019.2901468
Schram A (2005) Artificiality: the tension between internal and external validity in economic experiments. Journal of Economic Methodology 12(2):225–237
Schram A, Ule A (2019) Handbook of Research Methods and Applications in Experimental Economics. Edward Elgar Publishing. https://doi.org/10.4337/9781788110563
Schröter I, Krüger J, Siegmund J, Leich T (2017) Comprehending studies on program comprehension. In: International Conference on Program Comprehension (ICPC), IEEE, pp 308–311. https://doi.org/10.1109/icpc.2017.9
Scoccia GL, Malavolta I, Autili M, Salle AD, Inverardi P (2021) Enhancing trustability of android applications via user-centric flexible permissions. IEEE Trans Software Eng 47(10):2032–2051. https://doi.org/10.1109/TSE.2019.2941936
Shakeel Y, Krüger J, von Nostitz-Wallwitz I, Lausberger C, Durand GC, Saake G, Leich T (2018) (Automated) literature analysis - threats and experiences. In: International Workshop on Software Engineering for Science (SE4Science), ACM, pp 20–27. https://doi.org/10.1145/3194747.3194748
Sharafi Z, Huang Y, Leach K, Weimer W (2021) Toward an objective measure of developers' cognitive activities. ACM Transactions on Software Engineering and Methodology 30(3):30:1–30:40. https://doi.org/10.1145/3434643
Shargabi AA, Aljunid SA, Annamalai M, Zin AM (2020) Performing tasks can improve program comprehension mental model of novice developers: an empirical approach. In: International Conference on Program Comprehension (ICPC), ACM, pp 263–273. https://doi.org/10.1145/3387904.3389277
Shaw AD, Horton JJ, Chen DL (2011) Designing incentives for inexpert human raters. In: Conference on Computer Supported Cooperative Work (CSCW), ACM, pp 275–284. https://doi.org/10.1145/1958824.1958865
Shen Q, Wu S, Zou Y, Zhu Z, Xie B (2020) From API to NLI: a new interface for library reuse. J Syst Softw 169. https://doi.org/10.1016/j.jss.2020.110728
Shen Q, Wu S, Zou Y, Xie B (2021) Comprehensive integration of API usage patterns. In: International Conference on Program Comprehension (ICPC), IEEE, pp 83–93. https://doi.org/10.1109/ICPC52881.2021.00017
Shull F, Singer J, Sjøberg DIK (2008) Guide to advanced empirical software engineering. Springer. https://doi.org/10.1007/978-1-84800-044-5
Siegmund J, Siegmund N, Apel S (2015) Views on internal and external validity in empirical software engineering. In: International Conference on Software Engineering (ICSE), IEEE, pp 9–19. https://doi.org/10.1109/icse.2015.24
Simmons E, Wilmot A (2004) Incentive payments on social surveys: a literature review. Survey Methodology Bulletin 53
Singer E, Couper MP (2008) Do incentives exert undue influence on survey participation? experimental evidence. J Empir Res Hum Res Ethics 3(3):49–56
Singer E, Ye C (2013) The use and effects of incentives in surveys. Ann Am Acad Pol Soc Sci 645(1):112–141
Singer E, van Hoewyk J, Maher MP (1998) Does the payment of incentives create expectation effects? Public Opinion Quarterly pp 152–164

Sjøberg DIK, Hannay JE, Hansen O, Kampenes VB, Karahasanović A, Liborg NK, Rekdal AC (2005) A survey of controlled experiments in software engineering. IEEE Trans Software Eng 31(9):733–753. https://doi.org/10.1109/tse.2005.97
Sjøberg DIK, Dybå T, Jørgensen M (2007) The future of empirical methods in software engineering research. In: Future of Software Engineering (FOSE), IEEE, pp 358–378
Smith VL (1982) Microeconomic systems as an experimental science. Am Econ Rev 72(5):923–955
Smith VL (1994) Economics in the laboratory. Journal of Economic Perspectives 8(1):113–131
Soltani M, Panichella A, van Deursen A (2020) Search-based crash reproduction and its impact on debugging. IEEE Trans Software Eng 46(12):1294–1317. https://doi.org/10.1109/TSE.2018.2877664
Spadini D, Çalikli G, Bacchelli A (2020) Primers or reminders? the effects of existing review comments on code review. In: International Conference on Software Engineering (ICSE), ACM, pp 1171–1182. https://doi.org/10.1145/3377811.3380385
Stapleton S, Gambhir Y, LeClair A, Eberhart Z, Weimer W, Leach K, Huang Y (2020) A human study of comprehension and code summarization. In: International Conference on Program Comprehension (ICPC), ACM, pp 2–13. https://doi.org/10.1145/3387904.3389258
Stol KJ, Fitzgerald B (2020) Guidelines for conducting software engineering research. In: Contemporary Empirical Methods in Software Engineering, Springer, pp 27–62. https://doi.org/10.1007/978-3-030-32489-6_2
Taipalus T (2020) The effects of database complexity on SQL query formulation. J Syst Softw 165. https://doi.org/10.1016/j.jss.2020.110576
Taipalus T, Grahn H, Ghanbari H (2021) Error messages in relational database management systems: a comparison of effectiveness, usefulness, and user confidence. J Syst Softw 181:111034. https://doi.org/10.1016/j.jss.2021.111034
Tan SH, Li Z (2020) Collaborative bug finding for android apps. In: International Conference on Software Engineering (ICSE), ACM, pp 1335–1347. https://doi.org/10.1145/3377811.3380349
Teixeira S, Agrizzi BA, Filho JGP, Rossetto S, Pereira ISA, Costa PD, Branco AF, Martinelli RR (2020) LAURA architecture: towards a simpler way of building situation-aware and business-aware IoT applications. J Syst Softw 161. https://doi.org/10.1016/j.jss.2019.110494
Thompson RF, Spencer WA (1966) Habituation: a model phenomenon for the study of neuronal substrates of behavior. Psychol Rev 73(1):16
Tosun A, Dieste O, Vegas S, Pfahl D, Rungi K, Juristo N (2021) Investigating the impact of development task on external quality in test-driven development: an industry experiment. IEEE Trans Software Eng 47(11):2438–2456. https://doi.org/10.1109/TSE.2019.2949811
Uddin G, Khomh F, Roy CK (2020) Mining API usage scenarios from stack overflow. Inf Softw Technol 122. https://doi.org/10.1016/j.infsof.2020.106277
Uddin G, Khomh F, Roy CK (2021) Automatic API usage scenario documentation from technical Q&A sites. ACM Transactions on Software Engineering and Methodology 30(3):31:1–31:45. https://doi.org/10.1145/3439769
Uesbeck PM, Peterson CS, Sharif B, Stefik A (2020) A randomized controlled trial on the effects of embedded computer language switching. In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), ACM, pp 410–420. https://doi.org/10.1145/3368089.3409701
Urbieta M, Antonelli L, Rossi G, do Prado Leite JCS, (2020) The impact of using a domain language for an agile requirements management. Inf Softw Technol 127. https://doi.org/10.1016/j.infsof.2020.106375
Valderas P, Torres V, Pelechano V (2020) A microservice composition approach based on the choreography of BPMN fragments. Inf Softw Technol 127. https://doi.org/10.1016/j.infsof.2020.106370
Vassallo C, Proksch S, Zemp T, Gall HC (2020) Every build you break: developer-oriented assistance for build failure resolution. Empir Softw Eng 25(3):2218–2257. https://doi.org/10.1007/s10664-019-09765-y
Veen Fv, Göritz AS, Sattler S (2016) Response effects of prenotification, prepaid cash, prepaid vouchers, and postpaid vouchers: an experimental comparison. Social Science Computer Review 34(3)
Vegas S, Riofrío P, Marcos E, Juristo N (2020) On (mis)perceptions of testing effectiveness: an empirical study. Empir Softw Eng 25(4):2844–2896. https://doi.org/10.1007/s10664-020-09805-y
Viticchié A, Regano L, Basile C, Torchiano M, Ceccato M, Tonella P (2020) Empirical assessment of the effort needed to attack programs protected with client/server code splitting. Empir Softw Eng 25(1):1–48. https://doi.org/10.1007/s10664-019-09738-1
Wang X, Sanders GL (2019) For money, and for fun: exploring the effects of gamification and financial incentives on motivating online review generation. In: Americas Conference on Information Systems (AMCIS), AIS

Wang Y, Zhang M (2020) Reducing implicit gender biases in software development: does intergroup contact theory work? In: Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), ACM, pp 580–592. https://doi.org/10.1145/3368089.3409762
Wasserstein RL, Lazar NA (2016) The ASA statement on p-values: context, process, and purpose. Am Stat 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108
Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond "p < 0.05". The American Statistician 73(sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913
Weber RA, Camerer CF (2006) "Behavioral experiments" in economics. Exp Econ 9(3):187–192. https://doi.org/10.1007/s10683-006-9121-5
Weimann J, Brosig-Koch J (2019) Methods in experimental economics. Springer. https://doi.org/10.1007/978-3-319-93363-4
Wiese ES, Rafferty AN, Moseke G (2021) Students' misunderstanding of the order of evaluation in conjoined conditions. In: International Conference on Program Comprehension (ICPC), IEEE, pp 476–484. https://doi.org/10.1109/ICPC52881.2021.00055
Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer. https://doi.org/10.1007/978-3-642-29044-2
Wyrich M, Preikschat A, Graziotin D, Wagner S (2021) The mind is a powerful place: how showing code comprehensibility metrics influences code understanding. In: International Conference on Software Engineering (ICSE), IEEE, pp 512–523. https://doi.org/10.1109/ICSE43902.2021.00055
Yates R, Power N, Buckley J (2020) Characterizing the transfer of program comprehension in onboarding: an information-push perspective. Empir Softw Eng 25(1):940–995. https://doi.org/10.1007/s10664-019-09741-6
Zieris F, Prechelt L (2020) Explaining pair programming session dynamics from knowledge gaps. In: International Conference on Software Engineering (ICSE), ACM, pp 421–432. https://doi.org/10.1145/3377811.3380925

Jacob Krüger is Assistant Professor for software engineering at Eindhoven University of Technology, The Netherlands. He previously worked at Harz University of Applied Sciences Wernigerode, Germany, Otto-von-Guericke University Magdeburg, Germany, and Ruhr-University Bochum, Germany—and visited Chalmers University of Technology | University of Gothenburg, Sweden, as well as the University of Toronto, Canada. His research is concerned with software evolution, focusing on the interplay of human cognition and software quality that arise in evolving software systems.

Gül Çalıklı is a lecturer (Assistant Professor) in Software Engineering with the School of Computing Science at the University of Glasgow, United Kingdom. She received her Ph.D. and master's degrees in Computer Engineering and bachelor's degree in Mechanical Engineering from Boğazici University in Istanbul, Turkey. Her research interests include human aspects in software engineering, program comprehension and empirical software engineering.

Dmitri Bershadskyy is associated researcher at the Chair for Economic Policy at the Otto-von-Guericke University Magdeburg, Germany. Previously, he worked as research assistant at the Leibniz Institute for Economic Research in Halle, Germany. His research focuses on experimental economics, behavioral economics, digitization, and communication.

Siegmar Otto is professor at the chair of Sustainable Development and Change at the University of Hohenheim, Germany. He is active in work and organizational psychology, human-machine interaction, and algorithm-based decision systems – all under a sustainability perspective.

Sarah Zabel is associated researcher at the department of Sustainable Development and Change at the University of Hohenheim, Germany. Formerly, she has worked at the chair of Personality and Social Psychology at the Otto-von-Guericke University Magdeburg, Germany. Her research focuses on algorithmic systems and their influence on human decision-making as well as biases in software development.

Robert Heyer is Junior-Professor at the Leibniz-Institut für Analytische Wissenschaften Dortmund, Germany, and Bielefeld University, Germany. Before, he worked at the Otto-von-Guericke Unviersity Magdebrug and the Max Planck Institute for Dynamics of Complex Technical Systems. His research focuses on bioinformatics and the engineering of corresponding analysis systems.

Authors and Affiliations

Jacob Krüger[1] · Gül Çalıklı[2] · Dmitri Bershadskyy[3] · Siegmar Otto[4] · Sarah Zabel[4] · Robert Heyer[5,6]

✉ Jacob Krüger
j.kruger@tue.nl

Gül Çalıklı
HandanGul.Calikli@glasgow.ac.uk

Dmitri Bershadskyy
dmitri.bershadskyy@ovgu.de

Siegmar Otto
siegmar.otto@uni-hohenheim.de

Sarah Zabel
sarah_zabel@uni-hohenheim.de

Robert Heyer
robert.heyer@isas.de

1 Eindhoven University of Technology, Eindhoven, The Netherlands

2 University of Glasgow, Glasgow, United Kingdom

3 Otto-von-Guericke University Magdeburg, Magdeburg, Germany

4 University of Hohenheim, Stuttgart, Germany

5 Leibniz-Institut für Analytische Wissenschaften, Dortmund, Germany

6 Bielefeld University, Bielefeld, Germany