universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Paying Taxes. And Attention: Tax Compliance Behavior and Data Quality in Different Samples

verfasst von / submitted by

Benedikt Wilke, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Univ.-Prof. Dr. Erich Kirchler
Mitbetreut von / Co-Supervisor:	Mag. Dr. Jerome Olsen

Table of Contents

Acknowledgement

First, I would like to express my gratitude to my supervisors, Dr. Jerome Olsen and Žiga Puklavec, for their continuous guidance, support, and feedback throughout the process of this thesis. They showed me the path and how to follow it.

Further, I would like to thank Prof. Dr. Erich Kirchler for his supervision and share of expertise and everyone in his seminar for their valuable input.

I am grateful for the funding by the Department of Occupational, Economic, and Social Psychology, which made my research possible.

A special thanks goes to Bernd for helping me find the way through the jungle of statistics.

My sincerest gratitude goes to my girlfriend and all my friends for always having my back whenever I needed them.

Finally, I would like to thank my family for their unconditional and loving support in all these past years that brought me up to this point in the first place.

Abstract

Research in industrial, organizational, and economic psychology more often than not relies on convenience samples, that is, researchers use sampling methods that allow for easy and cheap access to participants. These different methods may not only result in different sample characteristics but could themselves have a confounding effect on the behavior being researched. As tax compliance behavior has been shown to depend on various situational and personal factors, this study examined differences in tax compliance behavior in four convenience samples. Those included a snowball sample recruited via Facebook posts, two samples recruited through online platforms (Prolific and SurveyCircle), and a sample recruited on university campus (laboratory sample). Additionally, differences in data quality, measured as response times and through attention check questions, were examined. Participants (N = 703) played an income tax game that was incentivized in two of the samples. In the game they had to earn a performance-based income and pay taxes on it over six consecutive rounds. Results showed that sampling methods explained differences in relative tax compliance levels which seemed to be driven by sociodemographic and motivational sample characteristics. The Facebook sample showed the highest tax compliance levels whereas the laboratory sample showed the lowest compliance levels. Data quality differed between samples with the SurveyCircle sample showing the fastest response times and with slightly higher passing rates for the attention checks in the Facebook sample. Results are discussed in the light of implications for the design of studies with mixed sampling modes and for the design of attention check questions.

Keywords: tax compliance, data quality, convenience sample, attention check question

1 Introduction

The decision whether to pay taxes in pre-modern times was not so much a question of honesty and integrity but a question of your liege lord's men taking perhaps a tenth of what you earned, or all of it. Today's tax systems are more complex but tax paying is still an important pillar of our society. Investigating why people pay their taxes today is therefore highly relevant, especially due to the fact that these individual tax payments impact the financing of essential goods and institutions, which provide, for instance, healthcare and education (Gangl, 2019).

Researchers of a range of disciplines have tried to develop adequate models for tax compliance behavior for decades (Alm & Malézieux, 2020). Helping to achieve this goal many participants have filled out numerous questionnaires, have spoken about their views on taxes in interviews, and have simulated the act of tax paying in so-called tax games. These samples of participants were collected to gain insights into bigger populations, as it is common procedure in social sciences. This method also prescribes a critical look at sampling techniques, as inferential statements about a population based on a poorly collected sample are a risk not only to the credibility of scientists and science itself but of any science-based policy making. Effects found in research must be replicated in further studies before being accepted as established (Munafò et al., 2017). This replication should be made with samples of the same and samples with different sociodemographic characteristics and with various sampling methods to be sure that the found effect is in fact present in more than just one study. As research on tax psychology is used to design tax systems and this research often relies on convenience sampling, it is imperative to know about possible biases induced by the sampling method. As representative samples in social sciences are oftentimes difficult and expensive to obtain, researchers tend to recruit those participants instead who are convenient to reach. Convenience samples therefore play an important role in academic research. Without questioning the use of convenience sampling in general, or being able to do so, a comparison of convenience samples of different types was made.

Thus, the aim of this study is to examine different samples typical for research in industrial, organizational, and economic psychology for differences in tax compliance behavior and data quality. I want to shed light on differences in tax compliance behavior that may arise due to different sampling methods as well as differences in data quality that the methods may bring with them.

In the following, I begin with the explanation of tax compliance, then I describe the measurement of data quality and its impact on study results. This is followed by a description of convenience samples and the hypotheses.

1.1 Tax Compliance

The classical explanation of decision making in the tax context follows entirely the line of the rationalism of homo oeconomicus. The homo oeconomicus describes that people think and decide rationally, they weigh up the various influencing factors and then decide in favor of the outcome that maximizes their benefit. The influencing factors in this explanatory approach, that was focused on income tax, are the tax rate, the amount of one's income, the probability of being audited by the tax authority, and the magnitude of fines in case of detected tax evasion (Allingham & Sandmo, 1972). The theory described that compliance could be increased by increasing audit probabilities and fine rates. These factors are called *deterrence factors* or *policy factors* (Alm, 2019) because they are aimed at deterring taxpayers from noncompliance and are determined by policies. While various empirical studies confirmed these theoretical considerations to some extent (see Alm, 2019; Kirchler et al., 2008; and Muehlbacher & Kirchler, 2016 for good overviews), it became rather clear that other factors must also play a role in tax compliance decisions. Otherwise, given the in reality very low probability of being audited by the tax authorities (Alm, 2019), no one would pay their taxes. Researchers in the field of psychology therefore proposed models that included other, not exclusively economic, factors as explanations. Braithwaite (2003) postulated a model of motivational postures, with which taxpayers may confront tax authorities. Several of these postures, such as *commitment* or *resistance*, may play a role simultaneously. Kirchler et al. (2008) suggested trust in tax authorities and perceived power of tax authorities as the two means through which tax compliance could be achieved. Additional factors for tax compliance, that they explained in relation to their *slippery slope framework*, were tax knowledge, the degree of participation in tax-related decision processes, attitudes toward taxes, personal, social, and national norms, and the perceived fairness of the tax system. Recently, the influence of emotions on decision making in the tax context has also been studied (Enachescu et al., 2019, 2020).

With so many factors found to influence tax compliance, it can be difficult for researchers to draw conclusions about causal relationships between the postulated influencing variables and tax compliance. This is especially true when using field studies or data collected by tax authorities because covariates are hard to control for and access to data is not always

granted. Thus, laboratory experiments have been a popular tool since the early days of research in this field. The classical tax experiment consists of several rounds, in which participants receive a certain income, on which they then must pay taxes. In doing so, they can decide whether to truthfully report the amount of their income or to make false statements in order to evade taxes. These tax experiments have been criticized for only being able to replicate income tax and also for not being realistic for a large number of countries and people (Levitt & List, 2007) because the necessary income tax information is provided by third parties (usually the employer) or the tax is withheld outright (Alm, 2019). Additionally, experiments that make use of students as main source for their participants can be criticized for their low external validity as students have been shown to be less compliant than non-students (Alm & Malézieux., 2020; Choo et al., 2016). However, these limitations should not outweigh the utility of laboratory experiments with tax games since no other method allows for such causal inferences. For some research questions, participants without much experience with paying taxes in the real world might be even beneficial (Muehlbacher & Kirchler, 2016). Starting here, this study further contributes to our understanding of what needs to be considered when using tax experiments to measure tax compliance behavior as this is especially important to ensure external validity.

1.2 Data Quality

High data quality is important in order not to jeopardize the validity of scientific research. Study participants who give answers in a way that has nothing to do with the content of the items to be answered, for instance, can endanger this quality. There are various potential reasons for why these invalid answers are given. Besides intentional deception or socially desirable answers, inattention when filling out questionnaires is one of these reasons. Inattention itself can also have various reasons, including boredom or distractions in the participant's environment. However, the reasons for inattention or carelessness are arguably less important here than their potential impact on study results. If inattentive responses are not detected as such and filtered out, the observed relationships between variables may be attenuated by reduced scale reliability, thereby increasing the risk of Type II errors in hypothesis testing (Meade & Craig, 2012). But also the opposite, the inflation of correlations and thus the higher risk of Type I errors, has already been pointed out as a possible consequence of inattentive response behavior (Huang et al., 2015; Wood et al., 2017).

For the most part, estimates of the prevalence of inattentive or careless responding depend heavily on the particular definition of inattentive responding and the cut-offs applied

to distinguish “good” from “bad” data. Meade and Craig (2012) found 10%–12% of their undergraduate participants to have responded carelessly in some way. Huang et al. (2012) suggested the term *insufficient effort responding* (IER) to subsume all types of inattentive or careless responses and found 5%–11% of their participants to engage in such behavior. Fleischer et al. (2015) reported to have found 15%–20% of their participants in online surveys to be inattentive when answering.

When attempting to counter the effects of poor data quality due to IER, various approaches are discussed in the literature. DeSimone et al. (2015) divide the different possibilities into three categories: direct, archival, and statistical screening methods. They suggest that the different methods can each detect different types of inattentive responders. The direct screening techniques are self-report questions, *instructed items*, and *bogus items*. These are also called *attention check questions* (ACQs; Cheung et al., 2017), because they are questions that are specifically used in a questionnaire to test participants’ attention. The archival techniques include, for example, the longstring technique, where participants that answer with the same response option for a row of items are screened out, and taking the response time as an indicator for inattentive behavior, while the statistical methods include, for example, using the personal reliability coefficient or psychometric synonyms as indicators where correlations of responses to item pairs are tested (DeSimone et al., 2015).

Overall, ACQs in the form of instructed and bogus items, are the most frequently suggested methods to test for inattentive response behavior among participants (DeSimone et al., 2015; see also Cheung et al., 2017; Fleischer et al., 2015; Landers & Behrend, 2015; Lovett et al., 2018; Meade & Craig, 2012). The simplest way is to implement self-report questions asking participants whether they responded truthfully, read the instructions carefully, or even if they think their data should be used or dismissed. Although this can serve as an initial way to sort out poor quality data, it apparently suffers from transparency and social demand characteristics. Bogus items, or infrequency items, are questions of which it is assumed that with a high probability all participants will give the same answer (while other answers are given *infrequently).* A typical example is the statement “I have 17 fingers on my left hand”. Although this seems like a safe way to sort out every inattentive reader at first glance, some precautions must be taken when implementing this type of item. As is the case for many other item types, ambiguity can easily render these items useless when participants understood them in a different way than intended by the researcher. Instructed items on the other hand bypass ambiguity by containing direct instructions on which answer to give. An example could be “On this question please answer with *fully disagree*”. Everyone who did not

comply with the instruction is considered to have been inattentive, at least in the moment of responding to this question.

Response times are paradata, that is, auxiliary data that can grant insights on how a questionnaire is filled out. Clicking through the questions without giving them much thought or even reading them, was the most frequent IER behavior detected by a study using mouse movement tracking (Stieger & Reips, 2010). Assuming that instructions, questions, and other items need certain minimal times to be read and answered, participants for whom faster times are recorded can be sorted out because their answers cannot be linked to the content of the texts (DeSimone, et al., 2015). Wood et al. (2017) examined response time as an indicator for data quality in online samples and recommend including it in research. Although cut-offs can be difficult to justify, Huang et al. (2012) suggest that times under 2 s per item are unlikely.

When it comes to the question of which methods, or better, which combination of them should be implemented, DeSimone et al. (2015) stress the point that to use all of them or even a random selection is not a fruitful approach because some of these methods may not be appropriate for specific study designs. Therefore, researchers should consider which forms of IER are most likely to occur in the specific survey design and sample chosen.

As attention can vary greatly between people and situations, the group of people studied, and the situation participants find themselves in while completing a questionnaire could have an impact on attention. If so, the sampling method itself acts as a bias for collected data and its interpretation, which makes it relevant to examine the quality of the collected data.

1.3 Convenience Sampling

A theoretical assumption for the use of inferential statistics is that a given sample, from which a population is to be inferred, is representative for that population. This representativeness typically is achieved by drawing samples in a random or quasi-random manner. However, as this often involves time-consuming and costly procedures, common practice in industrial-organizational psychological research is far from this ideal (Landers & Behrend, 2015). Instead, samples are drawn as *convenience samples* and true random sampling can be considered an exception. Convenience samples are those samples to which a researcher has easy access; their availability makes them convenient.

As major types of convenience samples Landers and Behrend name college student, online panel, crowdsourced, organizational, and snowball samples. College students seeking course credits are readily available in academic research but their comparability to the general

population has been questioned in prior research (e.g., Behrend et al., 2011; Roulin, 2015). Online panels are platforms where volunteers sign up for participation in studies, usually being paid a small amount of money as compensation. Crowdsourcing platforms are similar to online panels in that participants sign up there to complete tasks in exchange for payment, but they are not limited to research. The most well-known example for these platforms, at least for US-based research, is Amazon's Mechanical Turk (MTurk; https://www.mturk.com/; Behrend et al., 2011; Mason & Suri, 2012). It has seen a rise in use by researchers as recruiting platform in the past decade (Palan & Schnitter, 2018) and often serves as a source for comparisons between online and offline samples (e.g., Hamby & Taylor, 2016; Hauser & Schwarz, 2016). There is an ongoing debate between researchers over the use of online, crowdsourced samples, as some claim a superiority of offline samples over their online counterparts, especially regarding data quality (e.g., P.D. Harms & DeSimone, 2015). Others make a case for online panels as a new source of participants for researchers that should not be overlooked (e.g., Landers & Behrend, 2015; Lowry et al., 2016).

In the case of organizational samples, participants are drawn from an organization at hand, often one to which the researcher is linked in some way. This is the most used sample type published in industrial-organizational psychology (Landers & Behrend, 2015). Similarly, snowball samples are recruited through personal connections of the researcher or certain networks that give access to specific target groups that are otherwise hard to reach. While generally more typical for qualitative research, this kind of sampling technique plays an important role in student research.

1.4 The Present Study

A general criticism on experiments on tax psychology is that they provide little external validity, especially when relying on student samples (Muehlbacher & Kirchler, 2016). However, when we acknowledge that their advantages can outweigh their disadvantages for certain research questions and when we accept for the moment that most research in this field relies on convenience samples instead of probabilistic samples another question arises: Are those samples drawn from the same population? That is, can results from a study that used, for example, a student sample, be directly compared to results from a study that used, for example, a sample recruited via an online recruiting platform? And when we keep in mind that the quality of collected data can impact our results, then we also must ask if these different convenience samples yield different levels of data quality. These questions are what drove this study because it becomes obvious that mixed sampling modes can pose a

problem to the accuracy of measurement and thereby to the validity of results. This is the case when respondents answer differently on different survey modes and these answers are then looked at as if coming from one subject pool (Fang et al., 2014).

For this study several convenience samples were therefore compared in terms of their tax compliance behavior and the data quality they provided. To collect data on tax compliance behavior, participants were to play a short tax game. As the survey was programmed online and was going to be filled out on a computer (or other electronic devices), inattentively "clicking-through" the questionnaire seemed the most likely form of IER to occur. Thus, I chose to examine response times, that could be collected automatically, and passing rates of attention checks, that I implemented in the questionnaire, as indicators for IER.

1.4.1 The Four Samples

I will now discuss which samples were used in this study. All samples were recruited by convenience sampling. The four samples were (a) a sample, for which participants were recruited via posts on my Facebook page and email contact; (b) a sample, for which participants were recruited on the recruiting platform Prolific; (c) a sample, for which participants were recruited on the recruiting platform SurveyCircle; and (d) a sample, for which participants were recruited in the university building and through the online Laboratory Administration System for Behavioral Science (Labs) of the University of Vienna.

The Facebook Sample. Although not typical for published quantitative research, samples drawn from personal contacts of the researchers, such as friends, colleagues, and acquaintances met in real life or online, play an important role in student research. On social media sites, such as Facebook, there are groups where members post their studies looking for participants. Often, this is done on the researcher's personal profile page, too, thereby reaching out to those persons directly who are connected with the researcher, or the person who posts the study, on the particular social media site. To represent this practice, in which potential participants are asked directly by the researcher to participate, such a sample was included in the present study.

The Prolific Sample. Founded in 2014, Prolific (https://www.prolific.co) is an online subject recruiting platform where researchers can find willing participants for their surveys. Participants, on the other hand, are payed for their time spent on completing the surveys. While Amazon's MTurk sees the most widespread use in and out of research (Palan & Schnitter, 2018), Prolific tries to position itself as an alternative platform especially aimed at and designed for research (see Peer et al., 2017). As Prolific is based in Great Britain, so is its

largest participant group, with the USA coming in second. The rest is internationally diverse, with just over 2000 members indicating German as their nationality (https://www.prolific.co/demographics). This was the relevant participant group for this study because the survey was presented only in German. As online recruiting becomes a more important resource for researchers, Prolific was included as a sample in this study to represent online recruiting platforms where participants earn money by participating.

The SurveyCircle Sample. SurveyCircle (https://www.surveycircle.com) is an online subject recruiting platform that was founded in 2016 and aims to provide researchers with an easy-to-access subject pool. The underlying principle here is, that participants are not paid for participation but rather earn points[1]. These are then assigned to the participants' own survey, which in turn lets the survey rise in the global scoreboard of the platform. The higher a survey is ranked on the scoreboard the more points participants earn by filling out the survey, which makes the survey more attractive to complete. Points can also be saved for later surveys or spent on other members' surveys. This quid pro quo approach is often seen in social media groups where members exchange the links to their respective surveys and is professionalized by the ranking system of SurveyCircle. While this is admittedly an innovative way to create a participant pool, the system might bring its own drawbacks with it. On other platforms like Amazon's MTurk or Prolific, certain mechanisms are installed that allow researchers to reject submissions of poor quality. This penalizes participants, as their invested time is lost as a result and in the case of Amazon's MTurk even lowers their rating, which in turn lowers their eligibility for further tasks, thereby giving them incentives to provide high quality data. This incentive is not present in the case of SurveyCircle, as it does not matter which questionnaires members fill out and how attentively they do it, which could make them prone to provide lower quality data. To better understand how a participant pool that is based on a quid pro quo principle rather than monetary compensation differs from other subject pools, SurveyCircle was included as the third sample of this study.

The Laboratory Sample. As is the case for other universities, too, the Faculty of Psychology at the University of Vienna has its own database where students and other interested persons can register to participate in studies. This database, called Laboratory Administration for Behavioral Science (LABS), was the first source of participants for the laboratory sample of this study. The second source was recruiting in person and via announcements in the university buildings. The survey itself was then conducted in the social sciences laboratory where participants filled out the questionnaire at computers. While the

[1] There is, however, the option for researchers to incentivize participants through lotteries.

questionnaire itself was still provided online, the setting and recruiting process of this sample make it the only offline sample of the present study. Although technically free for all that are hinted toward the survey in some form, the place and means of the recruiting process used for these samples usually determine a high percentage of students in them. Student samples are commonly used in experimental studies investigating tax compliance behavior (Choo et al., 2016). To account for this, a student sample is included in this study as well.

1.4.2 Hypotheses

For the Facebook sample, I expected the demand characteristics provided by the sampling method (as theoretically described before and to be further discussed in section 2.3) to have a positive impact on tax compliance behavior. Survey participants were shown to alter their compliance when hinted toward a choice preferred by the experimenter (Navarick, 2007) and members of social network sites were shown to have a strong sense of reciprocity (Xu et al., 2012). Although no direct clue was given in this study as to which behavior was favored by me as the experimenter, the assumption was that people might show higher tax compliance as a "favor" to me as this is generally the socially more desirable behavior. Because SurveyCircle promotes itself as a platform were the members help each other out, I expected this thinking of "doing a favor" to positively influence tax compliance behavior in this sample, too. As mentioned before, students have been found to be less compliant than non-students (Alm & Malézieux., 2020). I therefore expected the tax compliance in the laboratory sample to be lower than in the other samples as the laboratory sample was thought to represent mainly a student sample. In addition, monetary incentivization could tempt participants to understand the tax experiment more as a gamble rather than a tax paying situation and thereby decrease compliance. This leads to the following hypotheses:

H1a: Participants of the Facebook sample will be more compliant than those of the laboratory and the Prolific sample.

H1b: Participants of the SurveyCircle sample will be more compliant than those of the laboratory and the Prolific sample.

Assuming that personal Facebook contacts form a "social bubble" of people with less diversity regarding interests and attitudes (Nikolov et al., 2015), I expected less variance for tax compliance behavior in this sample than it should be found for the general population. As students are a more homogeneous group than the general population as well (Hite, 1988), I expected low variance in the laboratory sample, too. As theoretically everyone can register to

the recruiting platforms Prolific and SurveyCircle, those two samples should therefore show higher variances in their compliance. The following hypotheses depict those assumptions:

H2a: Participants of the Prolific sample will show more variance in their compliance than those of the laboratory and the Facebook sample.

H2b: Participants of the SurveyCircle sample will show more variance in their compliance than those of the laboratory and the Facebook sample.

Social desirability has been shown to lead to slower response times (Furnham et al., 2013). Keeping with the aforementioned assumption that participants of the Facebook sample and the SurveyCircle sample are more prone to the effects of social desirability via the demand characteristics, I expected the Facebook sample and the SurveyCircle sample to take more time to fill out the survey. Fang et al. (2014), however, argue that satisficing, thus, the tendency to settle with the first sufficing option instead of the optimal one, could explain inconsistent results in response times found between different sampling methods better than social desirability. They also found that social media contexts, as present in the Facebook sample, can decrease satisficing behavior thereby slowing response times. Hamby and Taylor (2016) found that monetary incentivization, which is a sample characteristic of the Prolific sample and the laboratory sample in this study, increases satisficing behavior. Thus, I hypothesized as follows:

H3a: Participants of the Facebook sample will show higher completion times than those of the laboratory and the Prolific sample.

H3b: Participants of the SurveyCircle sample will show higher completion times than those of the laboratory and the Prolific sample.

Although varying results have been found on differences in data quality between various online panel samples, there is the finding that within the subject pool of Amazon's MTurk non-naïveté to ACQs is increasing as participants learn what these items look like (Hauser & Schwarz, 2016; Peer et al., 2014). This development could very much show the same pattern in Prolific with increasing time of existence of the platform. Additionally, members of Prolific are incentivized to yield high quality data because researchers can reject submissions and thereby payment to the participants. Assuming that these reasons might even

cancel out any inattentiveness stemming from the faster response times, I predicted the following[2]:

H4a: Participants of the Facebook sample will show lower passing rates for ACQs than those of the Prolific sample.

H4b: Participants of the SurveyCircle sample will show lower passing rates for ACQs than those of the Prolific sample.

2 Method

2.1 Participants

Using the software G*Power (Version 3.1.9.6; Faul et al., 2009), I calculated a necessary total sample size of 492 participants to detect an effect size of f = 0.15 with an alpha level of 0.05 and a test power of 0.8. I therefore took 123 as a minimal sample size per group. To protect against missing values and otherwise unusable data, as well as to be able to find even smaller effects, if there should be any, I took a group size of up to 200 participants as the intended optimal sample size. For different reasons, that are to be explained in section 2.3, I was not able to reach that number in each of the four groups. However, only the Facebook sample (n = 120) did not fully reach the minimal sample size.

For the present study data of a total 703 participants were used. Participants that did not complete the questionnaire up to at least the last page of the tax experiment were excluded[3]. See Table 1 for details of the sex, age, and nationality distribution over the different groups. As a priori exclusion criteria participants had to be at least 18 years old and fluent in German.

[2]An earlier draft of the study included the laboratory sample in this hypothesis, predicting higher passing rates for the laboratory sample than for the Facebook and the SurveyCircle sample. This was discarded when it became obvious that the ACQs could not be included in the laboratory sample questionnaire (see section 2.3).
[3] The certain number of thereby excluded participants is unknown because it was not possible to distinguish people that clicked on the link to the questionnaire multiple times from those that actually discontinued the questionnaire.

Table 1

Sociodemographic Information by Sample

Variable	Facebook	Prolific	SurveyCircle	Laboratory	Total
N	120	200	199	183	702
Gender[a]					
Female (%)	71 (59.2)	94 (47.0)	142 (71.4)	105 (57.1)	412 (58.6)
Male (%)	49 (40.8)	104 (52.0)	55 (27.6)	78 (42.4)	286 (40.7)
Mean age in years (*SD*)	37.94 (15.16)	30.83 (9.83)	26.90 (7.18)	23.62 (5.16)	29.05(10.55)
Students (%)	36 (30)	75 (37.5)	151 (75.9)	160 (87.4)	422 (60.1)
Place of Residence[b] (%)					
Germany	86 (71.7)	175 (87.5)	181 (91)	56 (30.6)	498 (70.9)
Austria	32 (26.7)	3 (1.5)	11 (5.5)	98 (53.6)	144 (20.5)

Note. Sociodemographic data were missing in one case in the laboratory sample. Participants of the laboratory sample were asked for their nationality instead of place of residence.
[a] Missing percentages are due to the response option "diverse".
[b] Missing percentages are due to the response options "Switzerland" and "other".

2.2 Materials

2.2.1 Experimental Design

The present study is a multiple-group comparison that borrows aspects of a quasi-experimental approach. It comprises a mixed design with one between-subject factor, that is, the sample (four levels: Facebook, Prolific, SurveyCircle, laboratory), and one within-subject factor, that is, the deterrence factors that come into play in the tax game. However, the independent variable for the hypotheses is sample affiliation alone, as no hypothesis relates directly to the within-subject variables (the deterrence factors). Those within-subject variables are audit probability, fine rate, and tax rate and have two levels each. The variation over the rounds of the tax game followed the already given pattern of the study of the laboratory sample to ensure comparability. The distribution of the parameters over the tax game rounds can be seen in Table 2. The dependent variables are the relative tax compliance, the passing rates of the attention check items, and the response times.

Table 2

Distribution of the Deterrence Factors Over the Tax Game Rounds

Round	Audit probability	Tax rate	Fine rate
1	1%	40%	0.5
2	15%	20%	0.5
3	1%	40%	1.5
4	15%	20%	1.5
5	1%	20%	0.5
6	15%	40%	0.5

2.2.2 Survey Design

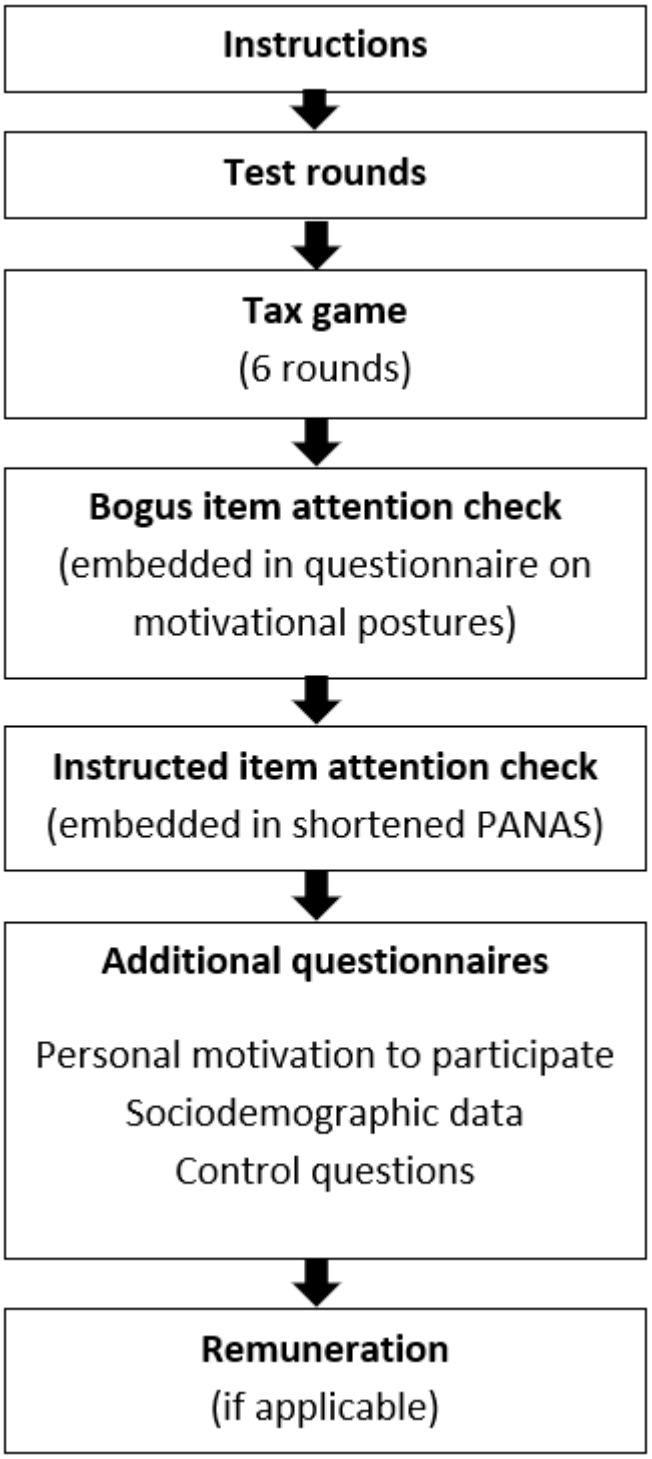
The survey was programmed using the online survey tool SoSci Survey (Leiner, 2019) and was made available via a link or, in the case of the laboratory sample, ready to start at a computer. After opening the link to the questionnaire and giving their consent to take part in the survey, participants read general instructions on the tax game and, in the case of the Prolific and the laboratory samples, received information about the determination of the payoff. They were then presented with three examples of a tax game round where they had to state their final net income, to check whether they understood the instructions. Afterwards, they were given a detailed depiction of the correct answers. Two test rounds followed, which resembled the actual game rounds with the exception that here audits did not occur randomly. Instead, in one test round participants were never audited, in the second, participants were always audited, to allow for a familiarization with both scenarios.

The tax game was at the core of the survey and in this case consisted of six rounds, preceded by the two test rounds. Each round participants earned a base income of 1000 Experimental Currency Units (ECU). In addition, each round started with a real-effort task where participants had to drag 10 sliders exactly on the 50%-mark within 20 seconds. For every successfully solved slider-task (Gill & Prowse, 2012) they earned 100 ECU for the round that were added to the base income. This let them earn a total income of up to 2000 ECU per round. Next, a tax had to be paid on this income and participants were informed of the exact amount of their income and the tax due (in percentage as well as absolute ECU). Additionally, they received information about the fine rate and the probability of an audit by tax authorities for the round. After participants entered the amount of taxes they wanted to pay, they immediately received feedback whether or not they had been audited and about the total net income of the round. If an audit had taken place and tax evasion was detected participants were informed about the amount of ECU that had been deducted from their income. After that, the next round started, and the process was repeated until six rounds had

been played. It is worth noting that for the laboratory sample the audit probabilities shown to participants were not true because audits were fixed to certain rounds, namely the second test round and the eighth round of the actual game. This resulted in participants of the laboratory sample only being audited in test round 2 because only data from the first six rounds were used for this study. After the last round of the tax game participants answered various questionnaires that partially included attention checks. Then, control questions asking for the parameters of the tax game were presented, and in the end, participants were debriefed and informed of the amount of their remuneration. An overview of the survey design is provided in Figure 1.

Figure 1

Flowchart of the Experimental Design



2.2.3 Measures

Tax Compliance Behavior. In the tax experiment tax compliance behavior was directly measured by the amount of taxes that participants declared. Thus, for each of the six rounds there was the amount of due tax and the amount paid. For further calculations of the tax compliance scores see section 2.4.1.

Attention/Insufficient Effort Responding. Several methods to detect IER were implemented in the study. However, this is only true for three of the four samples (Facebook, Prolific, and SurveyCircle), since data of the laboratory sample were taken from another study that had already been conducted when this study was planned. Thus, the following descriptions of measurements of IER apply only to the three samples, unless otherwise noted.

First, a bogus item was inserted into the questionnaire on motivational postures (see below). Participants had to indicate their agreement to the statement "The word *taxes* never appeared in the texts of this study" on a 5-point Likert scale (1 = *fully disagree,* 3 = *somewhat agree*, 5 = *fully agree*). Following the advice of Rouse (2015) and DeSimone et al. (2015) the item was created ad hoc and as an infrequency item. It also had a direct connection to the context of the study to better mask it and at the same time avoid prior knowledge of the item by participants with experience in filling out questionnaires. The rationale here was that it should be obvious to every attentive reader that they were asked to state their disagreement to the statement, because the word taxes had been presented in many occasions before within the tax game and its instructions. See a picture of the hidden bogus item in Appendix A "Page 40".

Second, an instructed item was implemented that asked participants to answer with a certain value on a 5-point Likert scale to demonstrate attention ("To demonstrate your attention, please enter the value *extremely* for the feeling *hostile*."). However, the instructions were put at the end of a somewhat lengthy introduction text to the questionnaire on the page concerned to better hide the item. This way only those participants that would read through the whole text (or at least take a closer look at it) were not going to miss the instruction. The questionnaire itself was a shortened version of the German version of the Positive and Negative Affect Schedule (PANAS; Krohne et al., 1996; Watson et al., 1988) that consisted of 10 adjectives with which participants were to describe their emotional state[4]. To pass this attention check participants were expected to accurately indicate feeling "extremely hostile", which I assumed hardly anyone would do for any other reason than being instructed to do so.

[4] I selected 10 (five of positive valence, and five of negative valence) with high reliability of the 20 adjectives of the PANAS. However, the apparent measurement of emotion served exclusively to mask the instructed item and none of the other items were used for any calculations.

See Appendix A “Page 41” for a picture of the hidden instructed item. To test hypothesis 4, the passing rates for the bogus item and the instructed item combined served as dependent variable.

Third, control questions were presented that asked participants to recall the correct probabilities of being audited that had been used in the tax game, the correct tax rates, and the correct amount of guaranteed base income that participants received every round. The questions were presented in a multiple-choice format. For the probabilities five options were provided (1%, 5%, 10%. 15%, 20%) of which two (1%, 15%) were correct. For the tax rates six options were provided (15%, 20%, 25%, 30%, 40% ,45%) of which again two (20%, 40%) were correct. For the base income five options were provided (500 ECU, 1000 ECU, 1300 ECU, 1500 ECU, 2000 ECU) of which only one (1000 ECU) was correct. For none of the items the number of correct answers was given and for each question it was possible to answer with none, some, or every option clicked. In the case of the laboratory sample a control question asked for the correct base income only, but instead of a multiple-choice format an open question was used. The control questions were put at the end of the study, after the socio-demographic measures and before the debriefing.

Fourth, an item asked participants of all four samples to indicate on a 5-point Likert scale how attentively they had read the instructions (1 = *not at all*, 5 = *very*; for the laboratory sample 1 = *not at all attentively*, 5 = *very attentively*).

Response times. Times spent on a page of a questionnaire were automatically saved for every participant and then served as dependent variable for hypothesis 4. See section 2.4.2 for further considerations made.

Personal motivation to participate. To help explain possible differences in tax compliance behavior and data quality between samples, a short questionnaire was included in the study that asked participants to rate on a 5-point Likert scale to what extend certain statements applied to them (1 = *fully disagree* , 5 = *fully agree*). The items were created ad hoc for the study and were intended to cover the most important motives one could have to participate in the study. As with the attention check questions and due to its origin in another study, the laboratory sample could not be taken into consideration for this measurement. To account for the different backgrounds of the sampling methods and platforms, the samples were partially presented with different items. A common stem of three items were presented to all samples concerned, namely: “I participated in this study because I find studies of this kind interesting”; “I participated in this study as a favor to the study director”; “I participated in this study because I want to support scientific research”. Additionally, for the Prolific

sample the item “I participated in this study because I earn money by it” was included and for the SurveyCircle sample the item “I participated in this study because it helps my own study gain more participants” was added.

Motivational postures and further measures. Two subscales on the motivational postures of commitment (e.g., “Paying tax is the right thing to do.”; eight items) and game playing (e.g., “I like the game of finding the grey area of tax law.”; five items) were presented in their German translations (Braithwaite, 2003; Kirchler & Wahl, 2010). Participants indicated their level of agreement to the statements on a 5-point Likert scale (1 = *fully disagree*, 5 = *fully agree*). Furthermore, sociodemographic data like age, gender, occupational status, and current residence were assessed and a question asking if participants had previously taken part in a study on tax behavior was included. With the exception of the sociodemographic information, none of the data described in this paragraph was collected in the laboratory sample.

2.3 Procedure

For the recruiting of the Facebook sample, I posted an announcement on my personal Facebook page where I asked any reader to take part in my survey and further share the link to the survey with friends, family members, or colleagues. After about a month I posted a second announcement, this time with a higher visibility than the first post due to an attached photograph, where I repeated my appeal. To further increase the reach, I sent emails to friends from whom I knew that they did not have a Facebook account or sent personal messages via instant messengers and posted my request in chat groups of which I was a member. To replicate an experimenter demand effect (Zizzo, 2010) that possibly comes with this very personal sampling technique and should therefore be expected to attenuate correlations between variables of interest in other studies using such samples as well, I specifically asked potential participants to do me the favor of participating and sharing the link to the study with others. Thus, any possible impact of a demand effect is part of the operationalization of this sampling method here as this is how requests to participate in studies are usually brought forward in social media. Recruiting took place during July 24, 2020 and August 26, 2020, when data of 120 participants were collected and after which no additional volunteer could be found within two weeks.

The Prolific sample was recruited via the online sampling platform Prolific. Participants were prescreened for their nationality to be German. Sampling began on September 9, 2020, at 11:19 am and ended at 1:49 pm of the same day when the set goal of

200 participants was reached. In this sample participants received a monetary compensation of £1.00 for their time, as this is a requirement by Prolific. Additionally, they were able to earn a bonus payment of up to £0.50 that depended on their decisions in the tax game. On average, they earned £1,33 (*SD* = 0.09; *min* = 1.13, *max* = 1.50). Financial expenses were covered by a grant of the Department of Occupational, Economic, and Social Psychology of the University of Vienna. Total expenses were £280.01, including a service fee for Prolific and taxes.

The SurveyCircle sample was recruited via the online sampling platform SurveyCircle. The ranking system for studies on SurveyCircle works via points that are earned by partaking in other members' studies, which in turn lets the researchers' own study rise in the ranking. The higher a study is ranked the more "incentive points" are earned by filling it out making it more interesting to complete for other members. Additionally, participants rate the study on a 5-star scale after completing it and this rating is shown in the overview alongside the number of raters. However, the study director can opt to hide this rating. Next to the rating is an approximate time needed to complete the study, as estimated by the study director. Participants indicate whether they think this time estimate is accurate, after completing the study, which, if true, is then indicated by a green clock symbol. This way members can determine which studies they want to prioritize for completing.

In general, there is a limit of 100 participants per researcher that can be recruited. On their website, SurveyCircle asks researchers interested in raising this limit to contact them directly. After having recruited 100 participants, I contacted SurveyCircle via email and was informed of their "pilot project", where they offered to raise the participant limit up to 200 for a donation of €40 (25 participants for each €10, payable all at once or in multiple steps). After the donation of €40, covered by the grant of the Department of Occupational, Economic, and Social Psychology, SurveyCircle additionally posted the link to the study on their social media channels and offered to raise the limit accordingly, if any cases of poor data quality should appear in the data.

I determined the threshold of rank 30 for my study to appear in the ranking without having to scroll down when using a standard computer monitor. This also was about the area where the first studies with more than 99 participants (indicated by "99+" next to the ranking) were situated. This way I had an estimate on how many points I had to earn to collect enough participants for my study. To reach this rank I filled out 73 studies, that earned me 1.16–10.00 points each (*M* = 5.25, *SD* = 1.56), prioritizing those with low time estimates and only taking

part in those for which I was eligible[5]. It is worth noting, that SurveyCircle does not prescreen participants and leaves it to study directors to implement screening questions in their questionnaires. However, there is the option to specify the target group in the description of the study so that honest participants can determine which studies to complete. For my study I put the minimal age of 18 years as exclusion criterion in the description.

Sampling began on July 25, 2020, and ended on August 12, 2020, when the limit of 200 participants set by the platform was reached. One participant was not counted correctly by the software, resulting in a total number of 199 participants.

The laboratory sample was recruited in the course of another survey. Data of that survey were used because of the closing of university premises due to the COVID-19 pandemic which made any data collection on site impossible at that point in time. The survey, of which data were taken instead, used the same tax experiment. Participants that had registered on the university's internal online platform Laboratory Administration for Behavioral Science (LABS) of the University of Vienna were informed of the study via email. Additionally, announcements were posted in university buildings and in various Facebook groups and active on-site recruitment was conducted and individuals were personally invited to participate in the experiment. Upon arrival at the laboratory, participants chose a seat at a computer where the link to the questionnaire was already opened. The experimenters instructed participants in a brief standardized introduction to read the instructions well and remain seated at the end of the experiment until the amount earned was paid out, after which participants started the survey.

Sampling took place on 9 days between November 25, 2019, and December 5, 2019. Of the 365 participants recruited for that survey I was able to examine only those 184 for the present study who had been assigned to the specific condition of the tax experiment that matched the one used in my study. As compensation for their time, participants of this sample received an average of €9.42 (*SD* = 1.97; *min* = 2.00, *max* = 12.00) that was comprised of a minimum payment of €2.00 and an additional bonus of up to €10 which depended on their decisions in the tax experiment.

[5]I do not have an exact time of how long it took me to complete these studies but 10 min per study is a good estimate, as most of the studies on SurveyCircle take around 5–15 min and I usually did not partake in studies above that time frame.

2.4 Data Preprocessing

2.4.1 Tax Compliance Behavior

Tax compliance behavior was operationalized by calculating the relative tax compliance ranging from 0 (= full tax evasion) to 1 (= full tax compliance). This value was the quotient of the amount of taxes paid in a round divided by the amount of taxes due to pay in that particular round. Thus, for each participant a score was calculated for each round.

A programming error in the code of the tax experiment made it possible for participants to pay a higher amount of taxes than they had to which resulted in relative compliance scores higher than 1. These overcompliant cases were set to 1 to facilitate calculations and interpretation of the results. Overcompliance was not equally distributed over the four samples, χ^2 (3) = 8.265, p = .041. According to the standardized residuals, participants of the laboratory sample were significantly less likely to be overcompliant than participants of the other samples (2 cases, z = -2.3).

To be able to test hypothesis 2, that predicted differences in the variation of relative tax compliance, I computed the mean tax compliance score and the standard deviation over all rounds of every participant. The coefficient of variation (CV) is the ratio of the standard deviation to the mean and can be interpreted as a measure of variation (Zöfel, 2003). However, it was not directly possible to compute this value for every participant, because some participants were completely noncompliant in all rounds, which resulted in a mean and standard deviation of zero. To avoid a division by zero, I excluded those cases of complete noncompliance, calculated the coefficient of variation for the remaining participants and then manually assigned a value of 0 to the completely noncompliant cases. Thereby I was able to still include those cases in the comparison while not changing the meaning of the coefficient of variation value as a case of complete compliance (M = 1, SD = 0, CV = 0) would have the same value as a case of complete noncompliance (M = 0, SD = 0, CV = 0) or any other case with a standard deviation of zero. It is worth mentioning that complete noncompliance was not equally distributed over the four samples , χ^2 (3) = 8.10, p = .044, with the adjusted standardized residuals indicating that participants of the Prolific sample were significantly more likely to be completely noncompliant than participants of the other samples (z = 2.4).

2.4.2 Response Times

The questionnaire programming software SoSci Survey automatically registers the time a participant spends on a certain page of the questionnaire. It also records the exact time the questionnaire was started and when it was ended. The times captured by SoSci Survey can

normally easily be taken to compute the response time for any participant. However, in this case it was not possible to directly compare these response times between the samples, because a different questionnaire was used for the laboratory sample, where participants played a longer tax game and were asked different questions after the tax game. To still be able to test hypothesis 4, I only counted those times together that were spent on pages that were comparable between questionnaires. The response times used for testing hypothesis 4 were therefore the sum of times spent on those pages of the tax game where participants indicated how much tax they wanted to pay and those pages where they received feedback on whether or not they were audited by the tax authority, including the two test rounds. In the case of the laboratory sample, only the times of the pages of the first six rounds (plus the two test rounds) of the tax game were taken into account.

An exclusion of outliers was necessary for this variable. I first excluded one extreme outlier in the Facebook sample who, according to the data, had spent more than 16 hours on one page of the questionnaire. Second, I excluded outliers that were more than three standard deviations away from the mean of the respective sample. This resulted in a total sample size of $N = 689$ for any calculation regarding the response times ($n = 117$ for the Facebook sample, $n = 195$ for Prolific, $n = 196$ for SurveyCircle, $n = 181$ for the laboratory sample).

2.4.3 Attention Check Questions

Dummy variables were created that indicated whether participants passed the attention check items or not. All were coded 0 = *attention check failed* and 1 = *attention check passed*. In the case of the instructed item, the check was considered passed only if the answer "extremely" was given. For the bogus item I followed the lenient way according to the recommendations of Kim (2018) and considered both the answers *disagree* und *fully disagree* as correct. Additionally, a dummy variable was created indicating whether both of the ACQs were passed or not, that is, participants that had responded correctly to both items were coded 1, all others were coded 0.

For the control questions, only those cases were coded as passed that answered with the exact combination of correct answers, when more than one answer of the multiple-choice options was right (e.g., the question for the correct tax rate hat two correct answers, namely, 20% and 40%).

3 Results

The results will be presented in different sections. First, I report descriptive statistics of the samples. Second, I describe the analyses of differences in relative tax compliance between the four samples. Third, I present the analyses on differences in the response times between the samples and in the passing rates of the attention checks. Last, I further describe differences between the samples that came to light but were not covered by the hypotheses. All analyses were computed using IBM SPSS Statistics (Version 24). The significance level was set at $\alpha = 5\%$ for all tests if not stated otherwise.

3.1 Descriptive Statistics

As a first overview, means and standard deviations are shown for various variables in Table 3. For the scores in the Tax Compliance column, I aggregated the relative tax compliances scores of the six tax game rounds to one score for each participant.

Table 3

Means and Standard Deviations (in parentheses) of Relative Tax Compliance, Response Times, and Motivational Postures (Game Playing and Commitment)

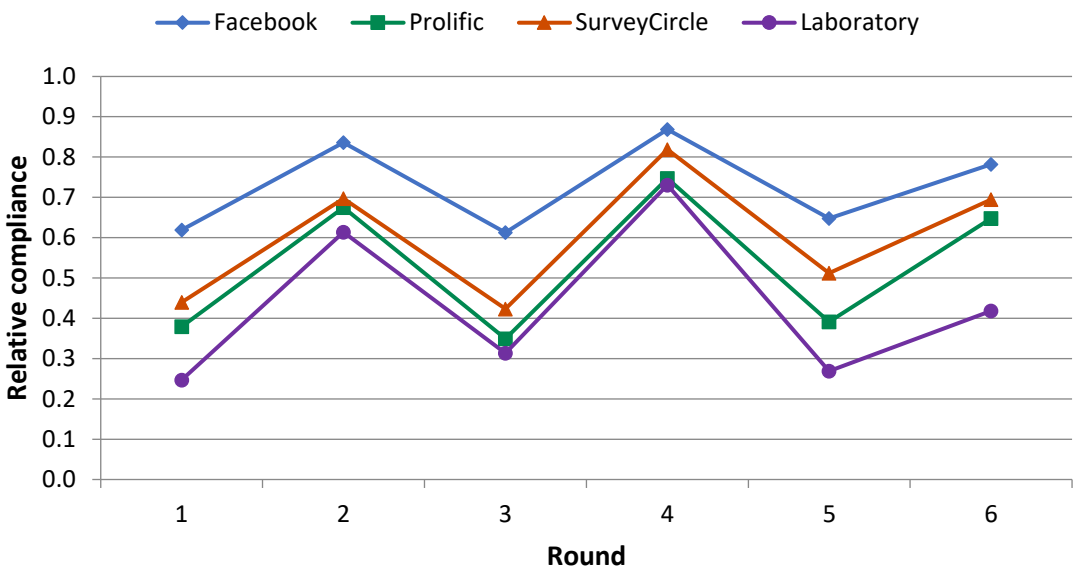
Measure	Facebook	Prolific	SurveyCircle	Laboratory
Tax compliance	0.73 (0.32)	0.53 (0.35)	0.60 (0.32)	0.43 (0.30)
Response time	191s (88)	170s (105)	142s (61)	151s (58)
Game playing	2.36 (0.81)	2.52 (0.86)	2.45 (0.92)	—
Commitment	3.91 (0.70)	3.84 (0.75)	3.79 (0.68)	—

Note. $N = 703$. The survey of which data for the laboratory sample was taken did not contain questionnaires on the two motivational posture scales; therefore data could not be obtained here, as indicated by dashes.

As displayed in Figure 2, relative tax compliance varied between the tax game rounds and between samples. See Table 2 in section 2.2.1 again for the distribution of the parameters of the tax experiment. These parameters determine the changes between rounds.

Figure 2

Relative Tax Compliance Across Rounds and Samples

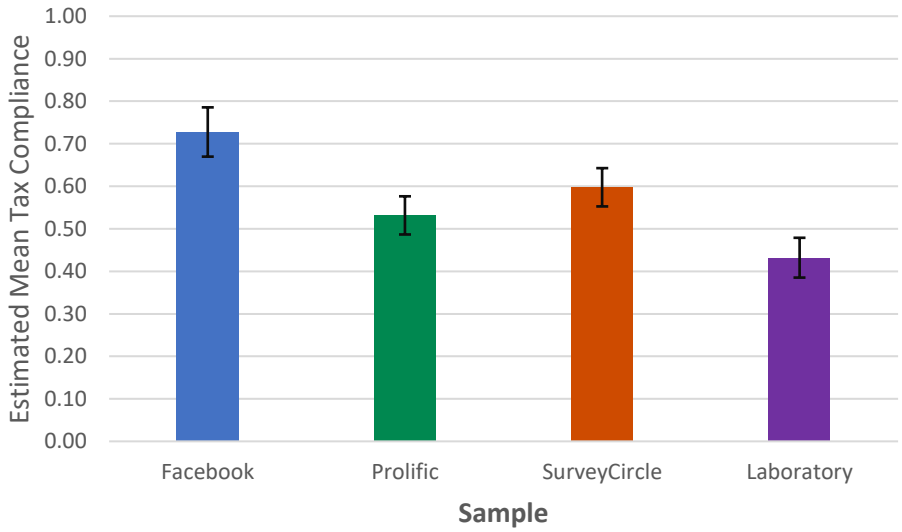


3.2 Relative Tax Compliance

I conducted a 2-way mixed analysis of variance (ANOVA) with round as the repeated measurement factor (within-subject factor) and sample as the between-subject factor. Mauchly's test for sphericity was significant, χ^2 (14) = 347.92, $p < .001$, which is why I used the Huynh-Feldt–corrected degrees of freedom for the tests of within-subjects effects (ε = .82). There was a significant main effect of round on relative tax compliance, F(4.10, 2862.66) = 206.26, $p < .001$, $\eta_p^2 = .23$, as well as a significant main effect of sample on relative tax compliance, F(3, 699) = 21.77, $p < .001$, $\eta_p^2 = .09$. This indicates that, ignoring sample affiliation, relative tax compliance scores differed across all rounds and that the samples differed in their relative tax compliance when the differences between rounds are ignored. The estimated mean for relative tax compliance of the Facebook sample was *estimated* $\mu = 0.73$, 95% confidence interval (CI) [0.67, 0.79], and thereby higher than the estimated mean of the Prolific sample, *estimated* $\mu = 0.53$, 95% CI [0.49, 0.58], and the estimated mean of the SurveyCircle sample, *estimated* $\mu = 0.60$, 95% CI [0.55, 0.64]. The lowest estimate stemmed from the laboratory sample, *estimated* $\mu = 0.43$, 95% CI [0.39, 0.48]. Those estimates are displayed in Figure 3.

Figure 3

Estimated Means of Relative Tax Compliance With 95% Confidence Intervals per Sample



Note. $N = 703$

Moreover, there was a significant interaction between round and sample, $F(12.29, 2862.66) = 5.14$, $p < .001$, $\eta_p^2 = .02$. This indicates that the profile of tax compliance scores, this is, the change of tax compliance scores from round to round, across all rounds was different for the four samples.

To further look for the differences indicated by the significant interaction, I ran multiple one-way ANOVA with sample as independent and the respective tax compliance scores of the six rounds as dependent variables. As Levene's test on homogeneity of variances was significant on a level of $p < .001$ for all rounds but round 3, I used Welch-statistics for the significance tests of the ANOVAs. All relevant values of the one-way ANOVAs are reported in Table 4. I used Games-Howell-tests for post-hoc comparisons to account for the violation of the assumption of homogeneity of variances and the unequal sample sizes. See Appendix B for an extensive look at comparisons between the samples over all six rounds. Overall, there was no clear picture of differences between the samples in the different rounds, but tendencies resembled the differences portrayed by the estimated means of the mixed ANOVA (Figure 3): Participants of the Facebook sample generally were more compliant than the other three samples as their scores differed significantly from those of all other samples in the first three rounds and from those of the Prolific and the laboratory sample in the second half of the tax game. Participants of the laboratory sample generally were less compliant than the other samples as their tax compliance scores differed significantly from those of all other samples

in rounds 1, 5, and 6. The Prolific and the SurveyCircle samples significantly differed in their relative tax compliance from one another only in round 5.

As the sample description suggested (see Table 1), gender was not equally distributed between the samples. Because four cells in the crosstab had expected values smaller than five, I ran Fisher's exact test, χ^2 (7) = 27.80, $p < .001$[6]. The standardized residuals indicated that in the Prolific sample there were significantly more male ($z = 2.5$), and less female participants ($z = -2.2$) than in the other samples, while in the SurveyCircle sample there were more female ($z = 2.3$) and less male participants ($z = -2.9$). As gender has been shown to have an influence on tax compliance behavior (Kastlunger et al., 2010), I conducted an analysis of covariance (ANCOVA) of relative tax compliance with gender as a covariate.

However, to be able to take gender as the covariate it had to be measured at a pseudo-metric level, that is, as a dichotomous variable. Therefore, I took the four participants that had indicated "diverse" as their gender and the one case of missing sociodemographic data out of the following calculations. I then ran the ANCOVA with sample and round as independent variables, relative tax compliance as dependent variable and gender as covariate. Again, the significant Mauchly-test for sphericity, χ^2 (14) = 330.372, $p < .001$, caused me to use the Huynh-Feldt–corrected degrees of freedom ($\varepsilon = .83$). The main effect of gender was significant, $F(1, 693) = 13.972$, $p < .001$, $\eta_p^2 = .03$, indicating that overall women were more compliant than men. The interaction between round and gender was significant as well, $F(4.14, 2869.90) = 4.09$, $p = .002$, $\eta_p^2 = .01$ meaning that women's and men's relative tax compliance varied differently in dependence to the tax game rounds. The main effect of round was still significant in this model, $F(5, 2869.90) = 10.67$, $p < .001$, $\eta_p^2 = .02$, as was the interaction effect of round and sample, $F(12.42, 2869.90) = 5.18$, $p < .001$, $\eta_p^2 = .02$.

[6] $N = 702$

Table 4

Means, Standard Deviations, and One-Way ANOVA Statistics for Relative Tax Compliance in All Tax Game Rounds

Round	Facebook		Prolific		SurveyCircle		Laboratory		ANOVA			
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F* ratio	*df*	*p*	ω^2
No. 1	0.62	0.44	0.38	0.44	0.44	0.44	0.25	0.38	20.00	3, 356.91	< .001	.26
No. 2	0.84	0.34	0.67	0.44	0.70	0.42	0.61	0.43	9.03	3, 370.59	< .001	.07
No. 3	0.61	0.43	0.35	0.44	0.42	0.44	0.31	0.41	13.56	3, 359.58	< .001	.10
No. 4	0.87	0.31	0.75	0.40	0.82	0.35	0.73	0.39	5.22	3, 368.53	= .002	.05
No. 5	0.65	0.45	0.39	0.47	0.51	0.47	0.27	0.41	20.89	3, 359.12	< .001	.13
No. 6	0.78	0.36	0.65	0.42	0.69	0.40	0.42	0.42	24.11	3, 364.83	< .001	.19

Note. N = 703. ANOVA = analysis of variance. Because of a violation of the assumption of homogeneity of variances in all but one ANOVAs, Welch-statistics are reported. ω^2 = effect size. See Appendix B for post-hoc comparisons of the samples over all six rounds.

The second hypothesis predicts higher variance in tax compliance scores for the Prolific and the SurveyCircle samples, compared respectively to the Facebook and laboratory samples. To compare the variation in relative tax compliance of the four samples, I conducted a one-way ANOVA with sample as the independent variable and the coefficient of variation as dependent variable. As the Levene's test of homogeneity of variances was significant, $F(3, 699) = 2.73$, $p = .043$, Welch-statistics were used. Differences between the samples were significant, $F(3, 364.75) = 14.69$, $p < .001$, $\omega^2 = .04$. I ran Games-Howell-tests for post-hoc comparisons as well as multiple t-tests. Table 5 displays the statistics of the six comparisons. Both methods revealed the same pattern: The Facebook sample ($M = 0.45$, $SD = 0.63$) showed significantly less variation than the Prolific sample ($M = 0.70$, $SD = 0.71$), than the SurveyCircle sample ($M = 0.69$, $SD = 0.69$), and the laboratory sample ($M = 0.98$, $SD = 0.72$). The Prolific sample and the SurveyCircle did not show significantly different variation, but both showed significantly less variation than the laboratory sample.

Table 5

Statistics of Pairwise Comparisons of the Coefficient of Variation Between Samples

Comparison	*df*	*t*	*p*[a]	*d*
FB vs. Pro	275.16[b]	−3.28	.001	−0.35
FB vs. SC	317.00	−3.09	.002	−0.35
FB vs. Lab	283.03[b]	−6.63	< .001	−0.70
Pro vs. SC	397.00	0.16	.875	—
Pro vs. Lab	375.49[b]	−3.71	< .001	−0.31
SC vs. Lab	381.00	−3.92	< .001	−0.31

Note. FB = Facebook; Pro = Prolific; SC = SurveyCircle; Lab = laboratory.

[a]The Bonferroni correction resulted in the new α-level of $p = .008$

[b]Equal variances not assumed according to Levene's test for equality of variances

3.3 Response time and attention checks

3.3.1 Response Time

As mentioned above, all calculations on response times were performed without the outliers and therefore with reduced sample sizes ($n = 117$ for the Facebook sample, $n = 195$ for Prolific, $n = 196$ for SurveyCircle, $n = 181$ for the laboratory sample). The computed response times in seconds were highly skewed and showed high kurtosis in their distributions. This caused me to run a Kruskal-Wallis-H test that revealed a significant difference between samples, $H(3) = 33.93$, $p < .001$. For direct comparisons, I conducted multiple Mann-Whitney tests and applied a Bonferroni correction which lead to a new level of significance of $p = .008$.

The Facebook sample (*Mdn* = 162, *IQR* = 85) had significantly higher response times than the Prolific sample (*Mdn* = 142, *IQR* = 81), $U = 8915.00$, $z = -3.23$, $p = .001$, $r = -.18$; than the SurveyCircle sample (*Mdn* = 130, *IQR* = 63), $U = 6972.50$, $z = -5.80$, $p < .001$, $r = -.33$; and than the laboratory sample (*Mdn* = 140, *IQR* = 69), $U = 7605.00$, $z = -4.11$, $p < .001$, $r = -.24$. The Prolific sample was significantly slower than the SurveyCircle sample, $U = 16058.50$, $z = -2.73$, $p = .006$, $r = -.14$; but did not differ from the laboratory sample, $U = 16646.50$, $z = -0.95$, $p = .342$, $r = -.05$. The SurveyCircle and the laboratory samples did not differ from one another in their response times either, $U = 15795.50$, $z = -1.84$, $p = .066$, $r = -.09$.

3.3.2 Attention Checks

For the determination of the passing rates of the ACQs, I compared the three samples, for which it was possible to collect the data, with regard to each of the two attention checks individually and then with regard to the two checks combined. First, I examined how many participants had passed the attention check of the instructed item. The proportion of participants who answered correctly differed significantly across samples, $\chi^2 (2) = 11.10$, $p = .004$. As can be seen in Table 6, the standardized residuals indicated that participants of the Facebook sample were significantly more likely to pass the attention check than participants of the other samples. The table also tells us that overall only 30.2 percent of participants passed the test. Pairwise comparisons with a Bonferroni correction of the α-level ($p = .017$) revealed that participants of the Facebook sample were 2.27 times more likely to pass the instructed attention check item than SurveyCircle participants, $\chi^2 (1) = 10.92$, $p = .001$, $OR = 2.27$, but did not differ significantly in their passing rates from Prolific participants, $\chi^2 (1) = 5.04$, $p = .025$, $OR = 1.72$. Participants from the Prolific and the SurveyCircle sample did not differ from one another either, $\chi^2 (1) = 1.47$, $p = .225$, $OR = 1.32$.

Table 6

Contingency Table for the Instructed Item per Sample

Sample	Instructed Item		
	Failed	Passed	Full sample
Facebook	68 (58.1%) −1.5	49 (41.9%) 2.3	117
Prolific	141 (70.5%) 0.1	59 (29.5%) −0.2	200
SurveyCircle	151 (75.9%) 1.0	48 (24.1%) −1.6	199
Total	360 (69.8%)	156 (30.2%)	516

Note. Cells depict the *n* of cases, percentages in the sample, and standardized residuals, if applicable.

Second, I examined how many participants had passed the attention check of the bogus item. Again the proportion of participants who answered correctly differed significantly across samples, χ^2 (2) = 12.46, p = .002. As displayed in Table 7, participants in the Prolific sample were significantly less likely to fail the attention check than participants of the other samples. In the SurveyCircle sample, in turn, participants were significantly more likely to fail the test. Overall, 90.3 percent of the participants of the three samples examined passed the bogus item attention check. Pairwise comparisons with the Bonferroni-corrected α-level of p = .017 revealed that participants of the Prolific sample were 3.17 times more likely to pass this attention check than participants of the Facebook sample, χ^2 (1) = 6.89, p = .009, OR = 3.17, and 3.95 times more likely to pass than participants from SurveyCircle, χ^2 (1) = 12.44, p < .001, OR = 3.95. Facebook and SurveyCircle participants did not differ from one another, χ^2 (1) = 0.40, p = .528, OR = 1.25.

Table 7

Contingency Table for the Bogus Item per Sample

Sample	Bogus Item		
	Failed	Passed	Full sample
Facebook	14 (11.7%) 0.7	106 (88.3%) −0.2	120
Prolific	8 (4.0%) −2.6	192 (96.0%) 0.8	200
SurveyCircle	28 (14.1%) 2.0	170 (85.9%) −0.7	198
Total	50 (9.7%)	468 (90.3%)	518

Note. Cells depict the *n* of cases, percentages in the sample, and standardized residuals, if applicable.

Last, I compared the proportion of participants that had passed both ACQs between the samples. In an overall comparison there were significant differences, χ^2 (2) = 10.22, p = .006, that seemed to stem from more participants in the Facebook sample that passed both attention checks, as indicated by the standardized residuals displayed in Table 8. Pairwise comparisons with a Bonferroni correction of the α-level (p = .017) neither showed differences in the passing rates between the Facebook and the Prolific sample, χ^2 (1) = 3.01, p = .083, *OR* = 1.53, nor between the Prolific and the SurveyCircle sample, χ^2 (1) = 2.79, p = .095, *OR* = 1.47. However, participants of the Facebook sample were 2.25 times more likely to pass both attention checks than participants of the SurveyCircle, χ^2 (1) = 10.24, p = .001, *OR* = 2.25.

Table 8

Contingency Table for Both Attention Checks Combined per Sample

Sample	Both Attention Checks combined		
	Failed	Passed	Full sample
Facebook	72 (61.5%) −1.3	45 (38.5%) 2.1	117
Prolific	142 (71.0%) −0.1	58 (29.0%) 0.2	200
SurveyCircle	155 (78.3%) 1.1	43 (21.7%) −1.8	198
Total	369 (71.7%)	146 (28.3%)	515

Note. Cells depict the *n* of cases, percentages in the sample, and standardized residuals, if applicable.

Furthermore, I conducted an exact McNemar test to examine the relation between the passing rates of the two attention checks. The McNemar test was to be preferred over a chi-square test in this case because the two dummy variables indicating the passing of the tests could not be considered independent. The test was significant, $p < .001$ (N = 515), telling us that the proportion of participants that had failed the attention checks differed between the two ACQs, that is, the bogus item attention check had a significantly higher passing rate than the instructed item attention check. This held true for each of the three samples.

3.4 Exploratory Analyses

In addition to the hypothesis testing, I compared the rates with which the control questions had been answered correctly. I also compared the scores of the motivational posture scales and the motivation to participate as indicated by the respective items.

3.4.1 Control Questions

The number of participants that correctly remembered the audit probability, the tax rates, and the base income used in the tax game varied between sample and item. Overall, only six participants answered correctly to all three questions and 22 participants did not answer correctly to at least one of the questions. Counts and percentages of responses to the questions per sample are displayed in Table 9.

I ran χ^2-tests to detect differences in the response rates between samples for the questions on audit probabilities and base income as well as Fisher's exact test on the question on tax rates due to expected frequencies smaller than five in two cells. The correct audit probabilities were remembered with significantly different rates, χ^2 (2) = 6.06, p = .048. The standardized residuals indicated that participants of the Prolific sample were more likely to remember the probabilities correctly than participants of the other samples (z = 2.3). Response rates for the question on tax rates differed significantly between samples, χ^2 (2) = 6.15, p = .047. The standardized residuals indicated that participants of the Prolific sample were less likely to answer correctly than participants of the other samples (z = −2.3). The correct base income was remembered with significantly different rates, too, χ^2 (2) = 6.97, p = .031, with the standardized residuals indicating that here participants from the Facebook sample were less likely to answer correctly than participants of the other samples (z = −2.6).

Table 9

Numbers of Responses to the Control Questions with Percentages per Sample

Sample	Audit probabilities		Tax rates		Base income	
	wrong	correct	wrong	correct	wrong	correct
Facebook	38	81	115	4	17	102
	31.9%	68.1%	96.6%	3.4%	14.3%	85.7%
Prolific	40	160	199	1	12	188
	20.0%	80.0%	99.5%	0.5%	6.0%	94.0%
SurveyCircle	54	145	191	8	15	184
	27.1%	72.9%	96.0%	4.0%	7.5%	92.5%
Total	132	386	505	13	44	474
	25.5%	74.5%	97.5%	2.5%	8.5%	91.5%

Note. Percentages are displayed for every item per sample. N = 518.

3.4.2 Personal Motivation to Participate

The means and standard deviations for the items on personal motivation to participate are displayed in Table 10. The main reason to participate for the Facebook sample and the

Prolific sample was to support scientific research. SurveyCircle participants indicated that their main reason was to gain participants themselves.

Table 10

Means and Standard Deviations of the Items on Motivation to Participate

Sample	Favor[a]		Research[a]		Interest[a]		Money[a]		Own study[a]	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Facebook	4.11	1.35	4.34	1.14	3.38	1.17	—	—	—	—
Prolific	2.40	1.40	4.36	0.87	4.25	0.99	4.20	0.93	—	—
SurveyCircle	2.81	1.53	3.98	1.06	3.34	1.13	—	—	4.06	1.28

Note. A dash indicates that the item was not provided for the respective sample.
[a]See section 2.2.3 and Appendix A "Page 42a" and "Page 42b" for the exact wording of the items.

3.4.3 Motivational Postures

Means and standard deviations of the two motivational postures scales are displayed in Table 2. After reviewing the data in P-P plots, I conducted Kruskal-Wallis-H tests because the assumption of normality was not met. This revealed no significant difference between the three samples for game playing, $H(2) = 2.72$, $p = .257$, nor for commitment, $H(2) = 5.06$, $p = .080$.

4 Discussion

The aim of this study was to examine different samples typical for research in industrial, organizational, and economic psychology for differences in tax compliance behavior and data quality. The different sampling methods with which the samples were recruited could lead to different sociodemographic characteristics of the samples and thereby to differences in the results of studies. They could also be a source of bias themselves as location and type of recruiting could have an influence on variables under examination. To contribute to insights into these relationships, a comparison was made between one sample recruited via Facebook posts, one sample recruited through Prolific, one sample recruited through SurveyCircle, and one sample recruited on university campus.

In doing so, I predicted higher relative tax compliance for the Facebook sample (H1a) and the SurveyCircle sample (H1b) compared to the Prolific and laboratory samples. However, the analysis of the collected data here provides support for only the first part of the assumption, as the SurveyCircle sample did not differ significantly from the Prolific sample. This suggests that the higher social proximity to the experimenter in the case of the Facebook sample may indeed have an impact on tax compliance behavior via demand effects. However,

this cannot be clearly separated from other effects that might arise from the sampling method. In the case of the SurveyCircle sample, the quid pro quo principle of the platform does not appear to have had an effect on tax compliance behavior as I predicted. Here, the anonymity of members seems to outweigh any positive effect of reciprocity on tax compliance behavior. In the case of the laboratory sample, the physical presence of the experimenter that could have had a positive influence on tax compliance behavior does not appear to do so as participants from this sample were the most dishonest.

The results of the exploratory analysis of the influence of gender on tax compliance behavior are in line with what other researchers found (e.g., Kastlunger et al., 2010). Women were more compliant than men in each of the rounds of the tax experiment. Interestingly, the gain in variation explained by gender seemed to come with a loss of variation explained by round while the variation explained by the interaction of round and sample stayed roughly the same, as indicated by the effect sizes of the ANOVA. This tells us, that the influence of the deterrence factors alone will be overestimated when gender is not controlled for.

I further predicted a higher variation of the tax compliance scores in the Prolific (H2a) and the SurveyCircle sample (H2b) compared to the Facebook and laboratory samples. This was only partly supported by the results, in that both samples varied more in their relative tax compliance than the Facebook sample but not more than the laboratory sample. In fact, the Facebook sample had the smallest variation of all four samples, suggesting it was the most intrinsically homogeneous sample when it comes to tax compliance behavior. This is in line with my reasoning that personal contacts on social media form a "social bubble" of people with little variation in interests and attitudes. Here, the population from which participants have been sampled seems to be rather homogeneous even when taking into account that this sample had the greatest variation in age of all four samples. However, in the end it cannot be exactly determined what is the actual cause for the smaller variation. The higher variation in the laboratory sample on the other hand suggests that participants on the Prolific and the SurveyCircle platforms are more similar to each other than participants that are mainly recruited on the university premises, at least when it comes to tax compliance behavior. However, again it cannot be said with certainty that this is to be attributed to the sampling methods.

To summarize the findings on tax compliance behavior, it can be said that differences in overall compliance, in variation of compliance within a sample, and in reaction to varying levels of deterrence factors can emerge between samples that are recruited from different populations. This falls in line with findings of differences in tax compliance behavior between

student and non-student samples (Alm & Malézieux., 2020; Choo et al., 2016), between different age groups, genders (Kastlunger et al., 2010), and the general finding that tax compliance behavior can vary very much between people and situations as there are many different motives to be compliant or non-compliant (Alm, 2019).

Interestingly, the use of monetary incentivization does not draw a clear line between the samples, as only one of the incentivized samples, the laboratory sample, differs from the other samples in tax compliance behavior, while the Prolific sample as the second monetarily incentivized sample shows the same relative tax compliance level as the SurveyCircle sample. This may be explained by the somewhat surprising finding that Prolific participants indicated money only as the third most important reason to participate.

Regarding the response times, I predicted that the Facebook participants (H3a) and the SurveyCircle participants (H3b) would be slower in filling out the questionnaire than their counterparts in the Prolific and the laboratory samples. I found support for only the first part of this assumption. Facebook participants indeed showed the highest response times with small to medium effect sizes. Although the original rationale for the prediction, namely, that the social media context of the recruiting method of this sample would decrease satisficing behavior and therefore increase response times might still hold true to some extent, other factors also could have played a role. The age distribution in this sample showed a second peak for ages 50–60 that explains the higher mean age in comparison to the other samples. As higher age has been linked to higher response times (Furnham, Hyde, & Trickey, 2013; C. Harms et al., 2017), this could be an additional factor in explaining the response time differences between samples. Furthermore, it stands to question how many participants actually were recruited through my posts on Facebook. Considering again the age distribution, personal feedback from friends I asked to participate, and observed peaks in questionnaire reflux during the sampling period, I suggest the explanation that the Facebook sample is in fact not really the Facebook sample I originally planned but rather a snowball sample with varied recruiting channels. This in return would weaken the explanation of a social media context having an influence on response times.

This apparently wrong assumption of sample characteristics can be transferred to the SurveyCircle sample as well. Here, I might have underestimated participants' non-naïveté with surveys and the high pressure to fill out many surveys to push the own survey in the ranking system which should result in increased satisficing behavior, thereby lowering response times. This is supported by SurveyCircle participants indicating that their main motivation to fill out the survey was to help their own studies gain more participants.

Additionally, a weakness in measurement is that I was only able to compare the response times for tax game rounds instead of completion times of the full questionnaire. While this still is a valid way to tell if participants differed in the time they took to decide on their tax compliance, it cannot tell us if they actually spend enough time to carefully read the items in the other parts of the survey.

For the attention checks, I predicted higher passing rates for the Prolific sample when compared to the Facebook (H4a) and the SurveyCircle (H4b) sample. This was not supported by the results when looking at the passing rates for both items combined. While Facebook participants passed the attention checks with a higher rate than SurveyCircle participants, the assumption that Prolific participants are more attentive when filling out a questionnaire than the other samples does not hold true. However, the most surprising part of the findings probably is the high rate of 69.8%, accumulated through all three samples, with which participants generally did not pass the instructed item attention check. Apparently, the majority of people did not fully read the introduction text to the questionnaire on emotional states. As alarming as this in itself may be, it is noteworthy that the questionnaire had a rather standard layout and many participants with experience in filling-out surveys may have seen it as overly tedious to read such a lengthy instruction for a questionnaire as self-explanatory as this one. Therefore, I suggest that my instructed item was ill-designed as it probably marked too many participants falsely as inattentive. Participants that did not read the instruction to the end did not necessarily provide poor quality data for the emotion scale. I therefore propose using only the passing rates of the bogus item attention check when looking for differences between the samples. Here, the rates of 4%–14.1% of participants marked as inattentive are in line with findings in the literature (Fleischer et al., 2015; Huang et al., 2012; Meade & Craig, 2010). When only the bogus item is used, support can be found for the predictions that Prolific participants will pass the attention check more often than participants of the Facebook and the SurveyCircle sample. This might be explained by their experience with surveys and the incentives to yield high-quality data that are provided by Prolific's system. Also, the large amount of time required to make one's study rise in ranking on SurveyCircle could encourage participants of the platform to engage in satisficing behavior, thereby decreasing attentiveness.

Summarizing the findings on data quality and IER it can be said that the examined samples differ in data quality, albeit with small to medium effect sizes. Furthermore, it must be stressed that the detection of IER behavior depends very much on the measurement used. Although infrequency items that are created ad hoc usually are more prone to be

misunderstood (see Curran & Hauser, 2019, for an interesting investigation of ambiguous ACQs) than instructed items, this is not a guarantee that the latter cannot be wrongly used, too, as it likely happened in this study. Still, it is worrisome that longer instruction texts are not read by a majority of participants. Although in simple questionnaires like the PANAS it probably does not pose a major problem, it should be taken into account when designing items where participants are asked to read long texts as a way of creating a scenario that is to be imagined while responding to the questions.

4.1 Limitations and Future Research

My study yields interesting insights into effects that potentially arise due to different sampling methods. It comprises a multivariate approach to the examination of these differences, and data from over 700 participants were collected. What it inherently cannot provide is an answer to the question what exactly is the cause for the differences found between samples. There are various potential biases that can have an influence on participants' behavior regarding tax compliance or data quality. Most importantly, the sociodemographic characteristics that define a sample may have a greater impact than the sampling method itself. Their impact cannot be distinguished from one another by this study. Also, although the procedure of playing the tax game and filling out the questionnaires itself can be considered fairly similar for the samples, there might be differences arising, for example, in response times, due to the device used as it was possible to open the link to the survey on mobile devices. A study examining the effects of questionnaire completion using a mobile device found lower data quality as well as longer completion times for mobile device user compared to PC users (Struminskaya et al., 2015). While it therefore was not ideal for this study that it was possible to complete the survey on a mobile device, there was no better option than recommending using a computer with a mouse in the survey description. Data on the device used was collected only in the Prolific sample and could therefore not be controlled statistically.

Additionally, differences in behavior could arise by the different circumstances in which participants fill out the survey and it is hard to tell which of these have the greater impact. While outliers in response times can easily be controlled by screening out these cases, the question remains whether someone who works on the survey in a concentrated way but takes a break of several hours really produces data of poorer quality than someone who fills in the questionnaire in one sitting but is distracted by influences in his immediate environment. In this study these contextual factors could only be controlled in the laboratory sample and

were considered to be part of the operationalization of the independent variable of sample affiliation. Although it might be difficult to control for these environmental covariates in online questionnaires, future research could focus on better ways to assimilate online surveys to laboratory settings to not miss out on this big subject pool of online panels while at the same time not losing too much internal validity.

Furthermore, there remains the question of how great the influence of the deterrence factors really was. While the variation between the rounds seems to be linked to the variation of, especially, the audit probability, the percentages of correct answers in the control questions leave doubt about how much participants responded to the deterrence factors at all. This, in turn, is mitigated by the apparent difficulty of the used multiple-choice questions with similar response options and the placement of these questions on the very end of the survey. Participants might have very well taken the deterrence factors into account when filing their taxes but then forgotten about their exact values over filling out the questionnaires on motivational postures, personal motivation to participate and sociodemographic data. To account for this, future surveys with a focus on the influence of these deterrence factors should provide these control questions directly after the tax game.

Typically, in tax experiments that want to investigate the influence of deterrence factors on tax compliance behavior, more rounds are played to allow for every combination of the deterrence factors. This study used only six rounds to decrease the time needed to complete the whole survey and because the influence of deterrence factors was not its main interest. However, it should be considered when comparing the results to similar experiments.

While for this study monetary incentivization was, much like the environmental influences during the process of filling out the questionnaire, considered part of the operationalization of sample affiliation, its possible impact on tax compliance behavior and even data quality cannot be distinguished of other underlying factors. Although it was not the goal to do so, and it probably would make for a very different, if not unrealistic, sample to pay participants for their participation in the case of the Facebook sample and would not even be possible for SurveyCircle, it should be considered when comparing the results.

Future research could use methods different to those used in this study to counter poor quality data, use different set-ups of the tax game, or connect the various sampling methods with further concepts from the field of tax psychology. Furthermore, there should be a closer investigation of potential differences of the samples regarding personality and other characteristics that have a known impact on tax compliance behavior and data quality to get a better picture of who is being questioned by the researchers. Additionally, the samples used in

this study could be compared to other convenience samples from other recruiting platforms such as Amazon's MTurk as they often give access to different populations. Perhaps most interestingly, a comparison of convenience samples and representative probabilistic samples should be made in order to allow for a better understanding of what can be and cannot be inferred from such surveys.

4.2 Conclusion

The present study investigated how different sampling methods can yield different results in a tax experiment and with regard to data quality. It compared four convenience samples in a multivariate design in terms of their tax compliance behavior, their response times, and their passing rates in attention check questions. Differences between the samples in tax compliance behavior and its variance came to light. This underscores the importance of interpreting results with the sample in mind that was used to infer the general population.

Furthermore, differences in data quality have been found that highlight the importance of including some kind of measurement of data quality in surveys. As was demonstrated by one attention check question used in the study, a majority of study participants do not read longer item explanations which bears the implication for researchers to keep item texts short.

Based on exploratory analyses, the observation made in this study is that younger people are less compliant in a tax experiment than older people, that women are more compliant than men, and participants who are monetarily incentivized tend to be less compliant although here no definite distinction can be made. Further, researchers should consider what the motivation for participants to take part in a study is likely to be as this seems to have an influence on what their response behavior looks like when filling out questionnaires.

Overall, this study shows how important the replication of findings is for the scientific method before these findings can be taken as a valid basis for decisions, especially in policy making. Finding an effect in one study does not implicate that it can be found again when sampling methods differ.

5 References

Allingham, M. G., & Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, *1*, 323–383. https://doi.org/10.1016/0047-2727(72) 90010-2

Alm, J. (2019). What Motivates Tax Compliance? *Journal of Economic Surveys*, *33*(2), 353–388. https://doi.org/10.1111/joes.12272

Alm, J., & Malézieux, A. (2020). 40 years of tax evasion games: a meta-analysis. *Experimental Economics*. https://doi.org/10.1007/s10683-020-09679-3

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. https://doi.org/10.3758/s13428-011-0081-0

Braithwaite, V. (2003). Dancing with tax authorities: Motivational postures and non-compliant actions. In V. Braithwaite (Ed.), *Taxing Democracy: Understanding Tax Avoidance and Evasion* (pp. 15–39). Aldershot: Ashgate.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations. *Journal of Business and Psychology*, *32*(4), 347–361. https://doi.org/10.1007/s10869-016-9458-5

Choo, C. Y. L., Fonseca, M. A., & Myles, G. D. (2016). Do students behave like real taxpayers in the lab? Evidence from a real effort tax compliance experiment. *Journal of Economic Behavior and Organization*, *124*, 102–114. https://doi.org/10.1016/j.jebo.2015.09.015

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. https://doi.org/10.1002/job.1962

Enachescu, J., Olsen, J., Kogler, C., Zeelenberg, M., Breugelmans, S. M., & Kirchler, E. (2019). The role of emotions in tax compliance behavior: A mixed-methods approach. *Journal of Economic Psychology*, *74*(November), 102194. https://doi.org/10.1016/j.joep.2019.102194

Enachescu, J., Puklavec, Ž., Olsen, J., & Kirchler, E. (2020, December 15). Tax compliance is not fundamentally influenced by incidental emotions: An experiment. https://doi.org/10.31234/osf.io/ra6ms

Fang, J., Wen, C., & Prybutok, V. (2014). An assessment of equivalence between paper and social media surveys: The role of social desirability and satisficing. *Computers in Human Behavior*, *30*, 335–343. https://doi.org/10.1016/j.chb.2013.09.019

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive Responding in MTurk and Other Online Samples. *Industrial and Organizational Psychology*, *8*(02), 196–202. https://doi.org/10.1017/iop.2015.25

Furnham, A., Hyde, G., & Trickey, G. (2013). On-line questionnaire completion time and personality test scores. *Personality and Individual Differences*, *54*(6), 716–720. https://doi.org/10.1016/j.paid.2012.11.030

Gangl, K. (2019). Status quo and future research avenues of tax psychology. In K. Gangl & E. Kirchler (Eds.), *A Research Agenda for Economic Psychology* (pp. 184–198). Edward Elgar Publishing.

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *102*(1), 469–503. http://dx.doi.org/10.1257/aer.102.1.469

Hamby, T., & Taylor, W. (2016). Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples. *Educational and Psychological Measurement*, *76*(6), 912–932. https://doi.org/10.1177/0013164415627349

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk Workers Ahead—Fines Doubled. *Industrial and Organizational Psychology*, *8*(02), 183–190. https://doi.org/10.1017/iop.2015.23

Harms, C., Jackel, L., & Montag, C. (2017). Reliability and completion speed in online questionnaires under consideration of personality. *Personality and Individual Differences*, *111*, 281–290. https://doi.org/10.1016/j.paid.2017.02.015

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hite, P. A. (1988). An examination of the impact of subject selection on hypothetical and self-reported taxpayer noncompliance. *Journal of Economic Psychology*, *9*(4), 445–466. https://doi.org/10.1016/0167-4870(88)90013-X

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Kastlunger, B., Dressler, S. G., Kirchler, E., Mittone, L., & Voracek, M. (2010). Sex differences in tax compliance: Differentiating between demographic sex, gender-role orientation, and prenatal masculinization (2D:4D). *Journal of Economic Psychology*, *31*(4), 542–552. https://doi.org/10.1016/j.joep.2010.03.015

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A., & King, K. M. (2018). Detecting random responders with infrequency scales using an error-balancing threshold. *Behavior Research Methods*, *50*(5), 1960–1970. https://doi.org/10.3758/s13428-017-0964-9

Kirchler, E., Hoelzl, E., & Wahl, I. (2008). Enforced versus voluntary tax compliance: The "slippery slope" framework. *Journal of Economic Psychology*, *29*(2), 210–225. https://doi.org/10.1016/j.joep.2007.05.004

Kirchler, E., & Wahl, I. (2010). Tax compliance inventory TAX-I: Designing an inventory for surveys of tax compliance. *Journal of Economic Psychology*, *31*(3), 331–346. https://doi.org/10.1016/j.joep.2010.01.002

Krohne, H., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica*, *42*(2), 139–156. http://doi.org/10.1037/t49650-000

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142–164. https://doi.org/10.1017/iop.2015.13

Leiner, D. J. (2019). SoSci Survey (Version 3.1.06). Retrieved from https://www.soscisurvey.de

Levitt, S. D., & List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World. *Journal of Economic Perspectives*, *21*(2), 153–174.

Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. J. (2018). Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology*, *67*(2), 339–366. https://doi.org/10.1111/apps.12124

Lowry, P. B., D'Arcy, J., Hammer, B., & Moody, G. D. (2016). "Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *Journal of Strategic Information Systems*, *25*(3), 232–240. https://doi.org/10.1016/j.jsis.2016.06.002

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Muehlbacher, S., & Kirchler, E. (2016). Taxperiments: About the external validity of laboratory experiments in tax compliance research. *Die Betriebswirtschaft*, *76*(1), 7–19.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Navarick, D. J. (2007). Attenuation and Enhancement of Compliance with Experimental Demand Characteristics. *The Psychological Record*, *57*(4), 501–515.

Nikolov, D., Oliveira, D. F. M., Flammini, A., & Menczer, F. (2015). Measuring online social bubbles. *PeerJ Computer Science*, *2015*(12), 1–15. https://doi.org/10.7717/peerj-cs.38

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Journal of Experimental Social Psychology Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Roulin, N. (2015). Don't throw the baby out with the bathwater: Comparing data quality of crowdsourcing, online panels, and student samples. *Industrial and Organizational Psychology*, *8*(2), 190–196. https://doi.org/10.1017/iop.2015.24

Stieger, S., & Reips, U. D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*(6), 1488–1495. https://doi.org/10.1016/j.chb.2010.05.013

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality. Evidence from a probability-based general population panel. *Methods, Data, Analyses*, *9*(2), 261–292. https://doi.org/10.4232/1.12245.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://doi.org/10.1037/0022-3514.54.6.1063

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response Speed and Response Consistency as Mutually Validating Indicators of Data Quality in Online Samples. *Social Psychological and Personality Science*, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168

Xu, C., Ryan, S., Prybutok, V., & Wen, C. (2012). It is not for fun: An examination of social network site usage. *Information and Management*, *49*(5), 210–217. https://doi.org/10.1016/j.im.2012.05.001

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98. https://doi.org/10.1007/s10683-009-9230-z

Zöfel, P. (2003). *Statistik für Psychologen (Im Klartext)*. Pearson Studium.

6 Appendices

6.1 Appendix A

Survey Material (German questionnaire)

Page 1: Introduction

Liebe Teilnehmerin, lieber Teilnehmer,

vielen Dank für Ihre Teilnahme an dieser Studie zu finanziellen Entscheidungen.
In diesem - **etwa 10 Minuten** dauernden - Fragebogen beschreiben wir Situationen, in denen Sie sich vorstellen sollen, ein Einkommen zu verdienen, welches Sie versteuern.

Ihre Teilnahme ist freiwillig. Sie können zu jedem Zeitpunkt und ohne Angaben von Gründen abbrechen.
Die Daten werden anonym und vertraulich behandelt und ein Rückschluss auf Ihre Person ist nicht möglich.

Bitte lesen Sie die Anleitung zur Studie auf den folgenden Seiten aufmerksam durch. Am Ende des Fragebogens werden wir einige Fragen zum Verständnis der Studie stellen.

Weiter

Page 2: Form of Consent

Hiermit bestätige ich, dass ich die Informationen auf der vorherigen Seite gelesen habe.

Ich weiß, dass die Teilnahme freiwillig ist und ich das Recht habe die Studie zu jedem Zeitpunkt und ohne Angabe von Gründen abzubrechen.

Ich erlaube Ihnen mit den gewonnenen Daten zu arbeiten und diese anonymisiert für wissenschaftliche Zwecke zu veröffentlichen.

Hiermit bestätige ich, dass ich an dieser Studie teilnehmen möchte:

- ○ Ja, ich möchte an dieser Studie teilnehmen.
- ○ Nein, ich möchte nicht teilnehmen.

Weiter

Page 2b: Information on Compensation (Only Provided in the Prolific Sample)

Teilnahmevergütung

In dieser Studie arbeiten Sie mit **echtem Geld.** Das bedeutet, dass Sie – zusätzlich zu ihrem Verdienst über Prolific (**1 Pfund**) – Geld verdienen als direkte Konsequenz Ihrer Entscheidungen, die Sie während der Studie treffen.

Dafür wird am **Ende der Studie** eine der 6 Runden **zufällig ausgewählt** und **in Pfund ausgezahlt**. Die in der Studie verwendete Währung (ECU) hat einen Umrechnungsschlüssel von **1000 ECU = 0,25 Pfund.** Somit lassen sich zwischen **0 und 0,5 Pfund** zusätzlich verdienen.

Weiter

Page 3: Instructions for the Tax Experiment

Die Studie besteht aus insgesamt 6 Runden zu finanziellen Entscheidungen. Die folgende Situation ist in jeder Runde gegeben:

Sie erhalten jede Runde ein **Basiseinkommen von 1000 ECU (Experimental Currency Units)**. Darüber hinaus können Sie ein **zusätzliches Einkommen von bis zu 1000 ECU** erzielen. Dies hängt von Ihrer individuellen Leistung in einer einfachen Aufgabe ab. Daraus ergibt sich pro Runde ein **mögliches Gesamteinkommen von 2000 ECU.**

Auf das verdiente Einkommen fallen Steuern an. Der vorgeschriebene **Steuersatz variiert** von Runde zu Runde und wird Ihnen immer angezeigt. Es ist dabei Ihre Entscheidung, wie viel Steuern Sie tatsächlich zahlen.

In jeder Runde besteht eine **bestimmte Chance** von der Steuerbehörde **kontrolliert** zu werden. Auch diese **Chance varriiert** und wird Ihnen immer angezeigt. Sollten Sie kontrolliert werden und weniger als den vorgeschriebenen Steuersatz gezahlt haben, müssen Sie die **fehlende Steuer nachzahlen**. Zusätzlich fällt eine **Strafe** an, deren Höhe variiert und Ihnen immer angezeigt wird.

Weiter

Page 4: Instructions on the Slider-Task

Um das **zusätzliches Einkommen** von bis zu 1000 ECU zu verdienen, ist es notwendig jede Runde eine Aufgabe zu erfüllen.

Sie werden 10 Balken mit je einem Schieberegler sehen und haben 20 Sekunden für die Bearbeitung der Aufgabe Zeit. Ihre Aufgabe ist es den Schieberegler **genau in der Mitte (50%)** eines jeden Balkens zu platzieren. Für jeden korrekt platzierten Schieberegler erhalten Sie **100 ECU**.

Bitte beachten Sie, dass Sie nur dann ein zusätzliches Einkommen erhalten, wenn der Schieberegler **exakt** bei 50% platziert ist. Liegt der Schieberegler beispielsweise bei 49% statt bei 50%, erhalten Sie kein zusätzliches Einkommen.

Auf dieser Seite finden Sie einen exemplarischen Balken. Sobald sich der Schieberegler bei 50% befindet, klicken Sie bitte auf "*Weiter*".

Hinweis: Es ist nicht notwendig, den Schieberegler zu ziehen. Sie können ihn auch direkt per Mausklick platzieren.

Weiter

Page 5: Comprehension Questions

Bitte bearbeiten Sie die folgenden Aufgaben zur Überprüfung des Verständnisses:

Beispiel 1

Sie erhalten ein Basiseinkommen von 1000 ECU und verdienen zusätzlich 800 ECU. Damit haben Sie ein Gesamteinkommen von 1800 ECU. Der Steuersatz beträgt 40%. Bezogen auf Ihr Einkommen sind das 720 ECU. Sie zahlen den gesamten Betrag von 720 ECU und es findet keine Steuerprüfung statt.

Wie groß ist Ihr Gesamteinkommen nach dieser Runde?

Beispiel 2

Sie erhalten ein Basiseinkommen von 1000 ECU und verdienen zusätzlich 900 ECU. Damit haben Sie ein Gesamteinkommen von 1900 ECU. Der Steuersatz beträgt 40%. Bezogen auf Ihr Einkommen sind das 760 ECU. Sie zahlen nur 500 ECU der vorgeschriebenen Steuer und es findet keine Steuerprüfung statt.

Wie groß ist Ihr Gesamteinkommen nach dieser Runde?

Beispiel 3

Sie erhalten ein Basiseinkommen von 1000 ECU und verdienen zusätzlich 700 ECU. Damit haben Sie ein Gesamteinkommen von 1700 ECU. Der Steuersatz beträgt 40%. Bezogen auf Ihr Einkommen sind das 680 ECU. Sie zahlen nur 280 ECU der vorgeschriebenen Steuer und es findet eine Steuerprüfung statt. Damit müssen Sie die fehlenden 400 ECU plus das Anderthalbfache des Betrags (600 ECU) als Strafe nachzahlen.

Wie groß ist Ihr Gesamteinkommen nach dieser Runde?

Weiter

Page 6: Comprehension questions — Correct Answers

Die richtigen Antworten im Detail:

Beispiel 1

Basiseinkommen	**1000 ECU**
zusätzlicher Verdienst	**+ 800 ECU**
Gesamteinkommen	**= 1800 ECU**
gezahlte Steuer (Steuersatz 40 % = 720 ECU)	- 720 ECU
Nettoeinkommen	**= 1080 ECU**

Beispiel 2

Basiseinkommen	**1000 ECU**
zusätzlicher Verdienst	**+ 900 ECU**
Gesamteinkommen	**= 1900 ECU**
gezahlte Steuer (Steuersatz 40 % = 760 ECU)	- 500 ECU
Nettoeinkommen	**= 1400 ECU**

Beispiel 3

Basiseinkommen	**1000 ECU**
zusätzlicher Verdienst	**+ 700 ECU**
Gesamteinkommen	**= 1700 ECU**
gezahlte Steuer (Steuersatz 40 % = 680 ECU)	- 280 ECU
Nachzahlung	- 400 ECU
Strafe (1,5 x 400 ECU Nachzahlung = 600 ECU)	- 600 ECU
Nettoeinkommen	**= 420 ECU**

Weiter

Page 7: Start of the Test Rounds

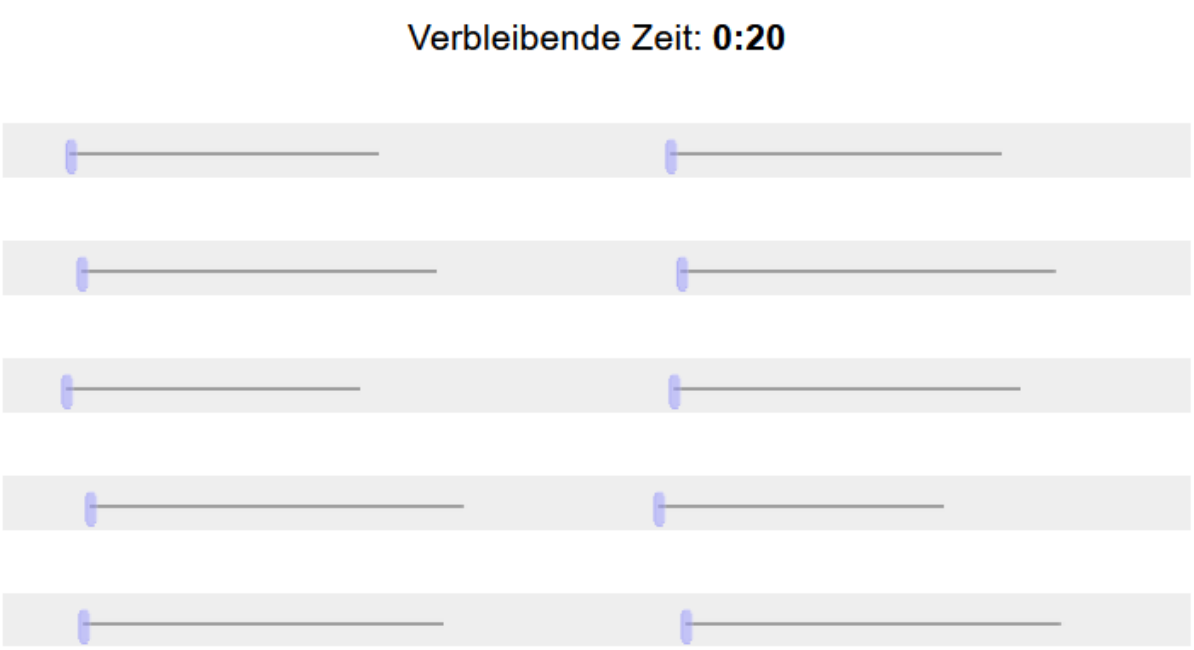
Im Folgenden finden zunächst **zwei Proberunden** statt.

Diese dienen lediglich dazu mit dem Ablauf vertraut zu werden.

Sobald Sie auf "*Weiter*" klicken, beginnen Sie mit der ersten Proberunde. Sie haben 20 Sekunden Zeit, um die Aufgabe zu lösen.

Weiter

Page 8: Slider-Task



Page 9: Overview of Earned Income and Tax Filing

Einnahmen in dieser Runde

Basiseinkommen	**1000 ECU**
Verdienst aus Aufgabe (7 gelöst)	**+ 700 ECU**
Gesamteinkommen	**= 1700 ECU**
anfällige Steuer (20%)	340 ECU

Ihre Einnahmen in dieser Runde betragen 1700 ECU.

Es ergibt sich eine Steuer von 340 ECU (20% Ihres Einkommens).
Falls Sie Steuern einbehalten und geprüft werden, müssen Sie die einbehaltene Steuer plus die Hälfte des Betrags der einbehaltenen Steuer als Strafe nachzahlen.

Die Wahrscheinlichkeit, von der Steuerbehörde geprüft zu werden beträgt für diese Runde 15%.

Bitte geben Sie an, wie viel Steuern Sie zahlen.

[ ] ECU

Weiter

Page 10: Feedback Phase When No Audit Occurred

Sie wurden **nicht geprüft**.

Einnahmen in dieser Runde nach Steuererklärung

Basiseinkommen:	**1000 ECU**
Verdienst aus Aufgabe (8 gelöst):	**+ 800 ECU**
Gesamteinnahmen:	**= 1800 ECU**
gezahlte Steuer (gefordert: 40% 720 ECU):	- 720 ECU
Gewinn:	**= 1080 ECU**

Weiter

Page 11: Start of the Second Test Round

Die erste Proberunde ist abgeschlossen.

Sobald Sie auf "weiter" klicken, beginnt die zweite Proberunde. Sie haben erneut 20 Sekunden Zeit, um die Aufgaben zu lösen.

Weiter

Pages 12–13 were identical to pages 8–9 but here an audit was forced

Page 14a: Feedback Phase When an Audit Occurred and No Tax Evasion Was Detected

Sie wurden **geprüft**.
Die Prüfung ergab, dass Sie die geforderten Steuern gezahlt haben.

Einnahmen in dieser Runde nach Steuererklärung

Basiseinkommen:	**1000 ECU**
Verdienst aus Aufgabe (7 gelöst):	**+ 700 ECU**
Gesamteinnahmen:	**= 1700 ECU**
gezahlte Steuer (gefordert: 20% 340 ECU):	- 340 ECU
Strafe:	- 0 ECU
Gewinn:	**= 1360 ECU**

Weiter

Page 14b: Feedback Phase When an Audit Occurred and Tax Evasion was Detected

Sie wurden **geprüft**.
Die Prüfung ergab, dass Sie die geforderten Steuern nicht vollständig gezahlt haben.

Einnahmen in dieser Runde nach Steuererklärung

Basiseinkommen:	**1000 ECU**
Verdienst aus Aufgabe (7 gelöst):	**+ 700 ECU**
Gesamteinnahmen:	**= 1700 ECU**
gezahlte Steuer (gefordert: 20% 340 ECU):	- 150 ECU
Nachzahlung:	- 190 ECU
Strafe:	- 285 ECU
Gewinn:	**= 1075 ECU**

Weiter

Page 15: Start of the Actual Six Tax Experiment Rounds

Die Proberunden sind nun abgeschlossen.

Bevor Sie mit der Studie weitermachen, befreien Sie sich bitte von jeglichen Gedanken und Gefühlen, die Sie gerade beschäftigen, und konzentrieren Sie sich auf die Situation, die Ihnen im Folgenden präsentiert wird. Bitte lesen Sie die folgenden Texte aufmerksam durch und konzentrieren Sie sich auf die jeweils beschriebene Situation.

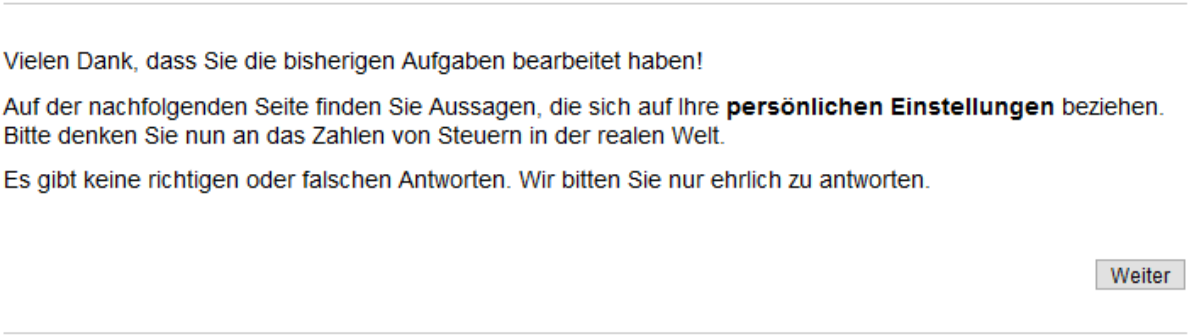
Im Anschluss an die Beschreibung beginnen Sie erneut mit der bekannten Aufgabe, für die Sie 20 Sekunden Zeit haben.

Klicken Sie auf "*Weiter*", wenn Sie bereit sind mit der **eigentlichen Studie** zu starten.

Weiter

Pages 16–38 were a repetition of the described procedure (i.e., slider-task, tax filing, feedback depending on occurrence of audit and detection of evasion)

Page 39: Transition Page to Questionnaires



Page 40: Questionnaire on Motivational Postures with Bogus Item

Bitte geben Sie an, in welchem Ausmaß Sie den folgenden Aussagen zustimmen.

	Trifft gar nicht zu		Trifft einigermaßen zu		Trifft völlig zu
Es gehört sich, seine Steuern zu zahlen	○	○	○	○	○
Ich überlege gerne welche Auswirkungen Veränderungen der Steuergesetzgebung auf mich haben könnten.	○	○	○	○	○
Steuern zu bezahlen ist eine Verantwortung, die von allen BürgerInnen gerne akzeptiert werden sollte.	○	○	○	○	○
Ich spreche gerne mit FreundInnen über die Lücken und Schlupflöcher im Steuersystem.	○	○	○	○	○
Ich fühle mich moralisch verpflichtet, meine Steuern zu bezahlen.	○	○	○	○	○
Es macht mir Spaß, die Lücken und Grauzonen des Steuerrechts herauszufinden.	○	○	○	○	○
Wenn ich meine Steuern bezahle, nützt das letztendlich Allen.	○	○	○	○	○
In den Texten dieser Studie kam das Wort *Steuern* nie vor.	○	○	○	○	○
Steuern zahlen hilft der Regierung sinnvolle Dinge zu tun.	○	○	○	○	○
Ich finde Vergnügen daran, einen Weg zu finden, wie ich meine Steuerzahlungen minimieren kann.	○	○	○	○	○
Alles in allem zahle ich gerne meine Steuern.	○	○	○	○	○
Die Steuerbehörde respektiert SteuerzahlerInnen, die sich nicht so leicht unterkriegen lassen.	○	○	○	○	○
Ich ärgere mich, meine Steuern zahlen zu müssen.	○	○	○	○	○
Ich sehe es als meine Verantwortung, meinen Steueranteil zu bezahlen.	○	○	○	○	○

Weiter

Page 41: Pseudo-PANAS with Instructed Item

Nun möchten wir gerne von Ihnen wissen, wie Sie sich während der Aufgabe fühlten.

Die folgenden Wörter beschreiben unterschiedliche Empfindungen. Lesen Sie jedes Wort und tragen Sie dann in die Skala daneben die Stärke Ihrer Empfindung ein. Sie haben die Möglichkeit, zwischen fünf Abstufungen zu wählen. Geben Sie bitte an, wie Sie sich während der Aufgabe fühlten. Um Ihre Aufmerksamkeit zu demonstrieren, geben Sie bitte beim Gefühl *feindselig* den Wert *äußerst* an.

	gar nicht		einigermaßen		äußerst
aktiv	○	○	○	○	○
beschämt	○	○	○	○	○
interessiert	○	○	○	○	○
feindselig	○	○	○	○	○
stolz	○	○	○	○	○
schuldig	○	○	○	○	○
wach	○	○	○	○	○
erschrocken	○	○	○	○	○
entschlossen	○	○	○	○	○
verärgert	○	○	○	○	○

Weiter

Page 42a: Questionnaire on Personal Motivation to Participate as Provided for the Prolific Sample

Bitte geben Sie an, in welchem Ausmaß Sie den folgenden Aussagen zustimmen.

Ich habe an dieser Studie teilgenommen,...	Trifft gar nicht zu		Trifft einigermaßen zu		Trifft völlig zu
... weil ich Studien dieser Art interessant finde.	○	○	○	○	○
... weil ich wissenschaftliche Forschung unterstützen möchte.	○	○	○	○	○
... um dem Studienleiter einen Gefallen zu tun.	○	○	○	○	○
... weil ich dadurch Geld verdiene.	○	○	○	○	○

Weiter

Page 42b: Questionnaire on Personal Motivation to Participate as Provided for the SurveyCircle Sample

Bitte geben Sie an, in welchem Ausmaß Sie den folgenden Aussagen zustimmen.

Ich habe an dieser Studie teilgenommen,...	Trifft gar nicht zu		Trifft einigermaßen zu		Trifft völlig zu
... weil ich Studien dieser Art interessant finde.	○	○	○	○	○
... weil ich wissenschaftliche Forschung unterstützen möchte.	○	○	○	○	○
... um dem Studienleiter einen Gefallen zu tun.	○	○	○	○	○
... weil dadurch meine eigenen Studie mehr Teilnehmer gewinnt.	○	○	○	○	○

Weiter

Page 43: Questions on Sociodemographic Information

Bitte geben Sie Ihr Alter an.

Bitte geben Sie Ihr Geschlecht an.

- ○ weiblich
- ○ männlich
- ○ divers

Bitte geben Sie Ihren derzeitigen Wohnsitz an.

- ○ Deutschland
- ○ Österreich
- ○ Schweiz
- ○ Anderes Land:

Was ist Ihr derzeitiger Beschäftigungsstatus?

Mehrfachantworten sind möglich

- ☐ Angestellte/r/Arbeiter/in
- ☐ Beamte/r
- ☐ Selbstständig
- ☐ Student/in
- ☐ Schüler/in
- ☐ In Pension
- ☐ Arbeitslos/Arbeitssuchend
- ☐ Sonstiges:

Weiter

Page 44a: Question on Subject of Studies (Only Provided When Occupation Student *Was Selected on Page 43)*

Bitte geben Sie Ihr Studienfach an.

Page 44b: Question on Prior Participation and Self-Report of Attentiveness

Haben Sie zuvor schon einmal an einer Studie zu Steuerverhalten teilgenommen?

- ○ Ja
- ○ Nein

Wie aufmerksam haben Sie die Instruktionen gelesen? gar nicht ○ ○ ○ ○ ○ sehr

Weiter

Page 45: Control Questions on Tax Experiment Parameters

Bitte denken Sie zurück an das Steuerspiel und beantworten Sie die folgenden Fragen!

Wie hoch waren die Wahrscheinlichkeiten, von der Steuerbehörde geprüft zu werden?

- ☐ 1%
- ☐ 5%
- ☐ 10%
- ☐ 15%
- ☐ 20%

Welche Steuersätze wurden über die Runden hinweg verwendet?

- ☐ 15%
- ☐ 20%
- ☐ 25%
- ☐ 30%
- ☐ 40%
- ☐ 45%

Wie hoch war das Basiseinkommen in allen Runden, welches garantiert war?

- ☐ 500 ECU
- ☐ 1000 ECU
- ☐ 1300 ECU
- ☐ 1500 ECU
- ☐ 2000 ECU

Weiter

Page 46: Debriefing (As Provided for the SurveyCircle Sample)

Herzlichen Dank für Ihre Teilnahme!

In dieser Studie untersuchen wir die Steuerehrlichkeit und die Aufmerksamkeit verschiedener Personengruppen bei der Beantwortung von Fragebögen.

Die präsentierten Abläufe des Steuerzahlens haben rein fiktiven Charakter.

Wichtig: Wir bitten Sie bis zum Ablauf der gesamten Studie am 31.07.2020 nicht mit anderen möglichen Teilnehmer*innen über die Studie zu sprechen, da wir weiterhin unvoreingenommene Teilnehmer*innen suchen. Vielen Dank!

Klicken Sie auf "*Weiter*", um die Studie abzuschließen. Auf der nächsten Seite finden Sie den Code, mit dem Sie Ihre Punkte auf SurveyCircle einlösen können.

Weiter

Page 47: Remuneration (Only Provided for the Prolific Sample)

Zur Auszahlung wurde **Runde 5** gewählt.

Ihr Nettoeinkommen in dieser Runde betrug **1600 ECU**.

Auf Basis des Umrechnungsschlüssels von 0,25 Pfund = 1000 ECU ergibt sich ein Bonus von **0.4 Pfund**. Dieser Bonus wird Ihnen in den kommenden Tagen ausgezahlt. Die Teilnahmevergütung in Höhe von 1 Pfund bekommen Sie wie gewöhnlich direkt über Prolific.

Bitte klicken Sie auf "Weiter" um auf Prolific weitergeleitet zu werden.

Wir danken Ihnen nochmals herzlich für Ihre Teilnahme.

Weiter

6.2 Appendix B

P-Values and Effect Sizes of Pairwise Post-hoc Comparisons (Games-Howell) of Compliance scores in Round 1

Sample	Facebook	Prolific	SurveyCircle	Laboratory
Facebook	—	−0.54	−0.40	−0.91
Prolific	< .001	—	0.14	−0.32
SurveyCircle	.003	.524	—	−0.46
Laboratory	< .001	.009	< .001	—

Note. P-Values of the comparisons are below the diagonal. Effect sizes (*Cohen's d*) are shown above the diagonal and have the direction of $sample_{column}$ to $sample_{row}$.

P-Values and Effect Sizes of Pairwise Post-hoc Comparisons (Games-Howell) of Compliance scores in Round 2

Sample	Facebook	Prolific	SurveyCircle	Laboratory
Facebook	—	−0.40	−0.36	−0.56
Prolific	.002	—	0.05	−0.14
SurveyCircle	.007	.953	—	−0.20
Laboratory	< .001	.512	.219	—

Note. P-Values of the comparisons are below the diagonal. Effect sizes (*Cohen's d*) are shown above the diagonal and have the direction of $sample_{column}$ to $sample_{row}$.

P-Values and Effect Sizes of Pairwise Post-hoc Comparisons (Games-Howell) of Compliance scores in Round 3

Sample	Facebook	Prolific	SurveyCircle	Laboratory
Facebook	—	−0.60	−0.44	−0.72
Prolific	< .001	—	0.17	0.26
SurveyCircle	.001	.335	—	−0.26
Laboratory	< .001	.844	.056	—

Note. P-Values of the comparisons are below the diagonal. Effect sizes (*Cohen's d*) are shown above the diagonal and have the direction of $sample_{column}$ to $sample_{row}$.

P-Values and Effect Sizes of Pairwise Post-hoc Comparisons (Games-Howell) of Compliance scores in Round 4

Sample	Facebook	Prolific	SurveyCircle	Laboratory
Facebook	—	−0.33	−0.15	−0.39
Prolific	.013	—	0.19	−0.04
SurveyCircle	.545	.229	—	−0.24
Laboratory	.004	.978	.094	—

Note. P-Values of the comparisons are below the diagonal. Effect sizes (*Cohen's d*) are shown above the diagonal and have the direction of $\text{sample}_{\text{column}}$ to $\text{sample}_{\text{row}}$.

P-Values and Effect Sizes of Pairwise Post-hoc Comparisons (Games-Howell) of Compliance scores in Round 5

Sample	Facebook	Prolific	SurveyCircle	Laboratory
Facebook	—	−0.55	−0.29	−0.88
Prolific	< .001	—	0.26	−0.28
SurveyCircle	.054	.049	—	−0.55
Laboratory	< .001	.036	< .001	—

Note. P-Values of the comparisons are below the diagonal. Effect sizes (*Cohen's d*) are shown above the diagonal and have the direction of $\text{sample}_{\text{column}}$ to $\text{sample}_{\text{row}}$.

P-Values and Effect Sizes of Pairwise Post-hoc Comparisons (Games-Howell) of Compliance scores in Round 6

Sample	Facebook	Prolific	SurveyCircle	Laboratory
Facebook	—	−0.34	−0.23	−0.91
Prolific	.014	—	0.12	−0.55
SurveyCircle	.188	.662	—	−0.67
Laboratory	< .001	< .001	< .001	—

Note. P-Values of the comparisons are below the diagonal. Effect sizes (*Cohen's d*) are shown above the diagonal and have the direction of $\text{sample}_{\text{column}}$ to $\text{sample}_{\text{row}}$.

7 Zusammenfassung

In der arbeits-, organisations- und wirtschaftspsychologischen Forschung wird häufig auf Convenience-Stichproben zurückgegriffen, d.h. die Forscher*innen verwenden Stichprobenmethoden, die einen einfachen und billigen Zugang zu den Teilnehmer*innen ermöglichen. Diese unterschiedlichen Methoden können nicht nur zu unterschiedlichen Stichprobencharakteristika führen, sondern könnten selbst einen Störfaktor für das untersuchte Verhalten darstellen. Da gezeigt wurde, dass Steuerehrlichkeit von verschiedenen situativen und persönlichen Faktoren abhängt, untersuchte diese Studie Unterschiede in der Steuerehrlichkeit zwischen vier Convenience-Stichproben. Diese bestanden aus einer Schneeballstichprobe, die über Facebook-Posts rekrutiert wurde, zwei über Online-Plattformen (Prolific und SurveyCircle) rekrutierte Stichproben und einer auf dem Universitätsgelände rekrutierten Stichprobe (Laborstichprobe). Zusätzlich wurden Unterschiede in der Datenqualität, erhoben als Antwortzeiten und durch Aufmerksamkeits-Check-Fragen, untersucht. Die Teilnehmer*innen (N = 703) spielten ein Einkommenssteuerspiel, das in zwei der Stichproben incentiviert war. Im Spiel mussten sie über sechs aufeinanderfolgende Runden ein leistungsabhängiges Einkommen verdienen und versteuern. Die Ergebnisse zeigten, dass die Stichprobenmethode Unterschiede in der Steuerehrlichkeit erklärte, was in soziodemografischen und motivationalen Stichprobenmerkmalen begründet zu liegen schien. Die Facebookstichprobe zeigte die höchste Steuerehrlichkeit, während die Laborstichprobe die geringste Steuerehrlichkeit zeigte. Die Datenqualität unterschied sich zwischen den Stichproben, wobei die SurveyCircle-Stichprobe die schnellsten Antwortzeiten aufwies und die Facebookstichprobe bei den Aufmerksamkeitskontrollen etwas höhere Bestehensquoten aufwies. Die Ergebnisse werden im Hinblick auf Implikationen für das Design von Umfragen mit gemischten Stichprobenmethoden und für das Design von Fragen zur Aufmerksamkeitskontrolle diskutiert.

Schlagworte: Steuerehrlichkeit, Datenqualität, Convenience-Stichproben, Aufmerksamkeitscheck