Journal of the Economic Science Association (2021) 7:167–183
https://doi.org/10.1007/s40881-021-00112-w

ORIGINAL PAPER

Running online experiments using web-conferencing software

Jiawei Li[1] · **Stephen Leider**[1] · **Damian Beil**[1] · **Izak Duenyas**[1]

Received: 5 January 2021 / Revised: 8 November 2021 / Accepted: 12 November 2021 /
Published online: 7 December 2021

Abstract

We report the results of a novel protocol for running online experiments using a combination of an online experimental platform in parallel with web-conferencing software in two formats—with and without subject webcams—to improve subjects' attention and engagement. We compare the results between our online sessions with the offline (lab) sessions of the same experiment. We find that both online formats lead to comparable subject characteristics and performance as the offline (lab) experiment. However, the webcam-on protocol has less noisy data, and hence better statistical power, than the protocol without a webcam. The webcam-on protocol can detect reasonable effect sizes with a comparable sample size as in the offline (lab) protocol.

Keywords Online experiment · Experimental methodology · ZTREE unleashed · Web-conferencing software · Webcam · Selection bias · Subject performance

JEL Classification C90 · C91 · D91

✉ Jiawei Li
jiawli@umich.edu

Stephen Leider
leider@umich.edu

Damian Beil
dbeil@umich.edu

Izak Duenyas
duenyas@umich.edu

[1] The University of Michigan, Ann Arbor, USA

1 Introduction

With the onset of the COVID-19 pandemic, experimental researchers have faced an increasing need to explore options for running online experiments—both to continue existing projects and to begin new projects. While existing platforms for running online experiments, such as MTurk, have been used successfully to study certain questions (Horton et al. 2011; Paolacci and Chandler 2014), other experiments may be impractical to run on those platforms due to task length, complexity, desired subject pool, or synchronous decision-making.[1] A natural goal, then, is to find an online experimental format that can approximate as closely as possible the typical offline (lab) experimental setting. To achieve this, we take advantage of another consequence of the pandemic: greater availability and subject comfort with web-conferencing software such as Zoom. Web-conferencing software has potential value for online experiments both in facilitating communication between experimenter and subjects, and to increase subject engagement and focus due to a more direct interaction with the experimenter (closer to what subjects experience in the lab). To explore the efficacy of this approach, we conduct experimental sessions using a combination of the newly developed ZTREE Unleashed (ZTu) (Duch et al. 2020) platform (which allows subjects to participate in an experiment programmed in Z-tree through their web browser) to run the experimental task, and a parallel Zoom session. We then compare both subject recruitment (demographics of participating subjects) and subject performance in the experimental tasks to sessions conducted in person shortly before the pandemic.

We explore two protocols for using web-conferencing software: whether subjects are asked to have their webcam on or off during the experiment. In both cases, the experimenter can use voice and text chat throughout the experiment to convey instructions and information, or to answer questions from subjects sent through text chats. Having the webcam off has the greatest level of subject anonymity and control (in terms of limiting the possibility that subjects use the webcam to influence one another); however, one may be concerned that subjects will be less engaged and focused on the task (early pilots suggested this may be a concern). Having the webcam on may help subjects stay on task and engaged, due to a feeling of being observed by the experimenter.[2] Our experiment uses an individual decision task,

[1] While MTurk has become a popular platform to recruit subjects and to organize online experiments, it has a number of design features that differ from the standard lab setting. MTurk recruits subjects from the internet instead of university students. While that has some benefits, it may not be ideal for experimenters that want to align with past lab studies as closely as possible. Additionally, MTurk is primarily used for surveys and relatively short tasks (often less than half an hour) (Arechar et al. 2018). Experimenters can also run into problems of high drop-out rates (Zhou and Fishbach 2016), and low subject effort (Chandler et al. 2014; Hauser et al. 2019).

[2] Our focus is on creating the feeling of the experimenter watching to make sure subjects stay on task that is typical in the lab context. One might also worry that the webcam could create or exacerbate experimenter demand effects. We do not think that is a concern in our setting. First, the primary task is a cognitively difficult dynamic decision problem that lacks a simple obvious "correct" answer. It is not clear what specific actions the subjects would think the experimenters "want." Second, if there was such a clear demanded behavior, it should distort the average choice/performance in Webcam versus No Web-

where subjects repeatedly solve a complex dynamic resource allocation problem. Therefore, we are not concerned about potential spillover between subjects from seeing others' reactions.[3] Our goal is to identify the experimental protocols (Webcam versus no Webcam) that have the best chance of generating comparable results as the offline (lab) experiment.

For comparability to lab studies, it is important that our protocol does not create distortions in subject recruitment or participation. For example, certain demographic groups (e.g., men versus women) may be more or less comfortable participating via web-conferencing software relative to the lab, which would create selection effects if those traits were also correlated with behavior in the decision task. We compare several demographic traits of participating subjects in our online formats to our lab data for the dynamic decision task. Neither online protocol leads to significant differences in subjects' age, gender or STEM major status based on the data from this dynamic decision-making study.[4] We also hope that subjects will perform the experimental task in a similar fashion to the lab format. We compare decision accuracy (with respect to the optimal policy) and profit earned in our primary dynamic decision problem between the online and lab data. We additionally look at three diagnostic measures: risk preferences, the cognitive reflection test (CRT), and the HIT-15 measure of backward induction ability. Overall, we find that subjects' performance in both online experimental protocols aligns with the lab experiment for both the primary task and the diagnostic measures. Using these protocols, we are able to generate high quality data using online experiments with complex and lengthy experimental tasks that require substantial subject attention and cognitive effort.[5]

Finally, we give some guidance about how experimenters may want to make design decisions such as sample size using these protocols. While both protocols yield data consistent with the lab data, we note that (1) the No Webcam protocol always leads to directionally worse average performance than the Webcam protocol, although the difference is not statistically significant; (2) in the primary dynamic decision problem, the No Webcam protocol leads to a larger variability in subject performance. This suggests that data with No Webcam is somewhat noisier, which may lead to issues if experimenters use too small a sample size. To examine this concern, using a simulation study, we find that both online experiment protocols have reasonable statistical power to detect the effect sizes observed in our data.

Footnote 2 (continued)
cam, e.g. by leading to improved performance in the most difficult version of the decision task. However, we see similar outcomes between protocols—just with less noisy data (explained in Sect. 3.4).

[3] This would be more of a concern for a strategic game. However, to address this, one could send subjects to breakout rooms either individually, or by role, and have separate experimenter connections to Zoom to monitor each room.

[4] As explained in Sect. 2.2, there are some subject characteristic differences between our online studies and another laboratory study we have demographic data on, but these differences are of comparable magnitude as between our lab study and that lab study. Hence, the observed selection effects there may simply reflect the natural variation between studies, rather than a systematic selection effect between online and offline studies.

[5] We note that some researchers comparing MTurk data from before and during the pandemic have seen subject performance degrade (Arechar and Rand 2021).

However, when we consider a range of possible effect sizes, the Webcam protocol is preferable. When effect sizes are moderate to large, the Webcam protocol has better power (due to lower performance variability) than the No Webcam protocol. For these treatment effect sizes, the statistical power of the Webcam protocol, and hence the required sample size, is comparable to the offline (lab) protocol.

2 The experiment

In this section, we give an overview of the experiment setup, subject recruitment, and the decision contexts we consider.

2.1 Experiment setup and subject recruitment

The online experiments were conducted using the newly developed platform ZTREE Unleashed (ZTu) that allows us to stream the ZTREE program to subjects who can then join the experiments remotely. The offline (lab) experiment was conducted in the Behavioral and Experimental Economics lab in the University of Michigan, Ann Arbor.

During the online experiments, we have a Zoom room set up in parallel to the experiment program.[6] The Zoom room mimics the lab experience of having all the subjects gather, check-in, conduct experiments together, and check out. We consider two protocols for the use of Zoom: requiring subjects to have their webcam on or off. To control for cross-subject communication and to protect subject anonymity, in both protocols: (1) subjects are only allowed to send private chats to the experimenter; the cross-subject chat function is disabled; (2) subjects are muted upon entry and throughout the experiment session; (3) subjects are renamed as soon as entering the Zoom room.[7]

Across the three formats, we utilize the same recruitment method using ORSEE (Greiner 2015), and recruit from the same subject pool. In the recruitment email for online experiments, subjects are notified that they will participate in the experiment by joining a Zoom room; they are also notified of the webcam-on requirement in the Webcam protocol, and we clearly noted that the video from the Zoom session would NOT be recorded. Subjects consist of undergraduate and graduate students from the University of Michigan, Ann Arbor. No subjects dropped out in any of the three

[6] Our protocol for using Zoom mirrors the procedures that Zhao et al. (2020) have used for their lab. We also thank Alex Brown and Valon Vitaku for helpful information about the approach to using Zoom monitoring for online experiments that they developed in parallel. Their preliminary data using a coordination game (played against past subject choices) also finds similar outcomes between lab and online format. We note that, to our knowledge, there is no existing study that compares different online experiment protocols (with and without webcam).

[7] The supplementary material offers additional details about the administration process we use, including figure illustrations for ZTu and Zoom setups.

formats.[8] A total of 194 subjects participated in our experiment: 69 subjects in No Webcam, 70 in Webcam, and 55 in the offline (lab) sessions.

2.2 The experiment decision tasks

We use the same decision tasks in all three experimental formats. Our primary decision task is a dynamic resource allocation problem modeled after a firm's product development process (described briefly below, additional details are provided in the supplementary material B). We also use three diagnostic tasks. Subjects' payments depend on their performance in all four tasks and range from 15 to 25 US dollars. The entire experiment lasts around 2 h for all three formats.

Dynamic resource allocation task

In the dynamic resource allocation task, subjects act in the role of a manager who is responsible for allocating a limited financial budget to sequentially arriving opportunities (abstractly representing potential design improvements). The manager must decide whether to accept the opportunity immediately and cannot wait and bundle the decisions. The opportunities vary in their benefits and costs and can only be accepted if enough budget is available. The manager's payoff is the total benefits accrued, plus any leftover budget.

Specifically, subjects begin with a budget of B experimental credits, which they can spend on up to 10 opportunities. Ex-ante, the benefits of the opportunities are random, drawn independently from a three-point distribution known to the subject. The task has two conditions (run between-subjects) that vary the cost of an opportunity. In the "simple condition", $B = 5000$, and the cost to implement each opportunity is fixed at 1000 credits throughout the project. In the "complex condition", $B = 6000$,[9] and the cost begins at 1000 but increases to 2000 for the last five opportunities (subjects know this in advance). This cost change substantially changes the optimal policy and makes the complex condition much more challenging for subjects. A session consists of subjects completing the allocation task 5 times under one cost condition (followed by the diagnostic tasks described below). We are interested in the gap in subject performance between the two conditions.

Diagnostic tasks and demographics

After the dynamic resource allocation task, subjects conduct three additional diagnostic tasks: (1) a risk preference measure, based on Holt and Laury (2002); (2) the cognitive reflection test (Frederick 2005);[10] (3) the Hit-15 task (Gneezy et al. 2010; Carpenter et al. 2013), which measures subjects' ability to backward induct.

[8] Two subjects in the Webcam protocol experienced internet connectivity issues during the session; however, they were able to complete the whole session, including the post-experiment demographic survey.

[9] The budget increase in the complex condition ensures that a comparable number of opportunities can be implemented between the two conditions.

[10] For the online sessions, we modified the wording of the CRT questions to prevent subjects from searching for the answers.

Table 1 Subject demographic information

Demographic information	Experiment format			Data from
	No webcam	Webcam	Lab	Study B
Proportion of female students	0.64	0.74	0.71	0.71
Average age	23.59 (5.23)	22.92 (4.26)	23.17 (5.78)	21.95 (3.09)
Proportion of STEM major students	0.58	0.56	0.49	0.39
Number of subjects	69	70	55	126

Standard deviations for age are in parentheses

Subjects have 5 minutes to complete each of the three diagnostic tasks, and they receive monetary payoffs based on their performance.

In the post-experiment questionnaire, we collect subjects' demographic information. Three kinds of information are collected: subjects' gender, age, and college major. Gender and age often play an important role in economic decision-making (Croson and Gneezy 2009; Kovalchik et al. 2005), and for our task, we anticipate that STEM majors may have better performance.

3 Experiment results

This section compares the results of the three experiment formats, with a focus on comparing subject recruitment (demographic characteristics of participating subjects) and subject performance.

3.1 Subject demographics

We compare the average age, gender, and share of STEM major students across each of our three experimental formats. As an additional point of reference, we also include the data from a *different unrelated study* (hereafter Study B) with 126 subjects. Both Study B and our offline (lab) experiment were conducted in the same lab in the same academic year and recruited from the same subject pool. In Table 1, we present the related results for the three experiment formats as well as the data from Study B.

When we consider the data from our dynamic decision-making problem, we find no indication that the experimental format influences any of the three demographic characteristics. None of the pair-wise comparisons between formats is statistically significant for either the age, gender ratio, or STEM ratio (proportions test for gender and STEM ratio, rank-sum test for age; $p > 0.20$ for all).

When using the data from Study B as the alternative lab benchmark, we first note that there is noticeable variability in age and proportion of STEM students across the two lab studies. With this in mind, we find that, compared to Study B, subjects in the No Webcam protocol are slightly older (23.59 vs. 21.95; rank-sum test $p = 0.01$);

Table 2 Subject demographic information—pooled

Demographic information	Pooled Online data	Pooled Lab data
Proportion of female students	0.69	0.71
Average age	23.26 (4.76)	22.32 (4.11)
Proportion of STEM major students	0.57	0.42
Number of subjects	139	181

Standard deviations for age are in parentheses

the other age and gender comparisons are not significant. We also find that both webcam protocols lead to significantly more STEM students (0.58 vs. 0.39, 0.56 vs. 0.39; proportions test $p < 0.03$ for both comparisons).

To further increase the power of our demographic comparisons, in Table 2, we present the results where we pooled the data from the two online formats (Webcam and No Webcam) and the two lab formats (the lab study of our dynamic decision problem and Study B), respectively, and compare the two pooled data. We find that the gender ratios are comparable between the pooled online and pooled lab data (0.69 vs. 0.71; proportions test $p > 0.20$). Meanwhile, subjects from the pooled online data are slightly older (23.26 vs. 22.32; rank-sum test $p = 0.06$), and significantly more likely to be in STEM majors (0.57 vs. 0.42; proportions test $p < 0.01$).[11]

In summary, based on the data from our dynamic decision-making experiment, both online experiment protocols can lead to generally comparable subject

[11] We conduct an ex-post power calculation by following the methods in (Howell 2012). The goal is to calculate the power given the sample sizes and conjectured mean differences between samples (formats); the results are all shown with respect to a significance level of 5%. For age analysis, the comparison between No Webcam (Webcam) and Study B leads to a power of 37% (47%) for a conjectured age difference of 1 year between formats. The pooled analysis does improve the power but not by much - the power becomes 52%. However, if we consider a conjectured age difference of 2 years, then we always get a good power: the power is higher than 90% for all the comparisons: No Webcam vs. Study B, Webcam vs. Study B, and Pooled Online vs. Pooled Lab. In other words, our current sample sizes are not well-powered to detect small age differences, but for any age differences larger than 2 years our samples sizes in either unpooled or pooled analysis are well-powered to detect them.

For the STEM ratio, for the two comparisons of No Webcam vs. Study B and Webcam vs. Study B, we derive a power of 52% for a conjectured proportion difference of 0.15. In the pooled analysis, however, we derive a power of 75% for a conjectured proportion difference of 0.15 between the two formats. Hence, the pooled analysis does help to significantly improve the power of our conclusions in STEM ratio comparisons.

Finally, for the ratio of female students, in our study the comparisons are not significant in any pair-wise comparisons across formats, either pooled or unpooled. Nonetheless, it may be useful to conduct an ex-post power calculation to determine the minimal “treatment effect" (the mean differences between two samples (formats)) such that it can be detected with a high power of 80%, given the sample sizes we have here. We find that such minimal treatment effect is 0.20 (0.19) when we consider No Webcam vs. Study B (Webcam vs. Study B), and is equal to 0.15 when we consider Pooled Online vs. Pooled Lab.

Table 3 Dynamic resource allocation decision task

	Experiment format		
	No webcam	Webcam	Lab
Simple condition			
Average decision accuracy rate	0.53 (0.05)	0.57 (0.03)	0.57 (0.04)
Average normalized profit ratio	0.79 (0.06)	0.81 (0.04)	0.82 (0.03)
Number of subjects	19	20	25
Complex condition			
Average decision accuracy rate	0.54 (0.03)	0.55 (0.03)	0.56 (0.03)
Average normalized profit ratio	0.26 (0.07)	0.29 (0.07)	0.38 (0.08)
Number of subjects	50	48	30

We have more subjects in the complex condition since we observe larger variability of profit performance there. Standard errors are in parentheses

demographic characteristics as the offline (lab) experiment. We observe subject characteristics differences between our online studies and another laboratory study (Study B); most noticeably, the proportion of STEM students are higher in our online experiments. However, the differences are comparable to the demographic differences between the two lab studies (our lab study of dynamic decision-making and Study B). Hence, such observed selection effect may be explained by the natural variation between studies rather than a systematic selection effect between online and offline (lab) studies.

3.2 Subject performance: dynamic resource allocation task

We next examine subjects' performance in the primary decision task.[12] Here, we compare subjects' payoffs to what can be achieved using the optimal policy (via dynamic programming). We consider two performance metrics: (1) decision accuracy rate; (2) normalized profit ratio. The decision accuracy rate is the percent of decisions coinciding with the optimal policy. For the normalized profit ratio, we begin with a benchmark profit achieved by the conservative policy of accepting the first 5 opportunities. The normalized profit ratio is then: (realized profit - benchmark profit)/(optimal policy profit - benchmark profit). Hence a profit ratio near 1 indicates the subject achieved nearly optimal profit, while a ratio near 0 indicates the subject did barely better than the benchmark. The results are summarized in Table 3.

For the simple condition, we observe that, directionally, the No Webcam protocol leads to the worst performance for both decision accuracy and profit among the three formats. However, none of the pair-wise comparisons in either performance metric is significant (rank-sum test $p > 0.20$ for all). The conclusion is the same in the complex condition. The No Webcam protocol leads to the lowest performance

[12] We remove the performance data from the two subjects who experienced continuous internet connectivity issues in the experiment.

for both metrics, but the difference is not significant ($p > 0.20$ for all pair-wise comparisons).

We are also interested in the difference in performance between the simple and complex conditions, fixing the experimental format.[13] We find that, regardless of the format, the following conclusions are true: (1) subjects have similar average decision accuracy rate between the simple and the complex condition (rank-sum tests $p > 0.50$); (2) subjects have substantially lower normalized profit ratios in the complex condition ($p < 0.01$). However, we note that the magnitude of the performance decrease from moving from simple to complex is largest in the online formats. The profit ratio decreases by 0.53 and 0.52 for the No Webcam and Webcam formats, while it only decreases by 0.44 in the lab format, although regressions suggest that the difference in magnitude of the treatment effect (i.e. the difference in profit ratios between the two conditions) is not significant.[14] One interpretation is that as we increase the complexity of the decision task, the potential concerns about subject attention and focus from online experiments have a larger impact on subject performance.

Next, we use regression analysis to study the robustness of our conclusion, particularly to the inclusion of demographic characteristics that may vary by format. We consider the following simple regression model that regresses the performance metric (either decision accuracy or normalized profit) on treatment dummy variables and demographic information.

$$\text{Metric}_i = \beta_0 + \beta_{\text{NW}}\text{NoWebcam}_i + \beta_{\text{W}}\text{Webcam}_i + \text{Age}_i + \text{STEM}_i + \text{Female}_i + \epsilon_i. \tag{1}$$

Here, i denotes the subject. Metric_i denotes the relevant performance metric (decision accuracy rate or normalized profit ratio). NoWebcam_i and Webcam_i are treatment dummy variables. STEM_i, and Female_i are demographic dummy variables. We run the regression for the two conditions (simple and complex) separately. The results are summarized in Table 4.

In both conditions (simple and complex) and for both metrics, we find that including the demographic information variables does *not* change our conclusion that both online experiment protocols are able to generate comparable results as the offline (lab) experiment (the p-value for the estimates of β_{NW} and β_{W} is always larger than 0.10 in Table 4).

3.3 Subject performance: diagnostic tasks

We now consider subject performance in the three diagnostic tasks: risk preferences, CRT and Hit-15. From Table 5, we see that the different formats yield comparable subject performance for all three tasks (rank-sum test $p > 0.15$ for all pair-wise

[13] This is the primary focus of our companion paper (available on request). In that paper, we explore the mechanisms underlying the performance difference summarized here, as well as potential interventions to improve performance in the complex condition.

[14] We pool the data from both conditions of the three protocols together, and regress the profit ratios on treatment dummies, online protocol dummy, and the dummy interaction terms: $\text{Profit Ratio}_i = \beta_0 + \beta_{\text{IC}} \cdot \text{Increasing Cost}_i + \beta_{\text{NW}}\text{No Webcam}_i + \beta_{\text{W}} \cdot \text{Webcam}_i + \beta_{\text{ICNW}}\text{Increasing Cost}_i \cdot \text{No Webcam}_i + \beta_{\text{ICW}} \text{Increasing Cost}_i \cdot \text{Webcam}_i + \epsilon_i$. Both interaction effects are insignificant (p value>0.40), hence we cannot reject the null hypothesis of equal treatment effects between the two online and the lab protocols.

Table 4 Dynamic resource allocation decision task regression analysis

	Estimates for the independent variables				
	No webcam	Webcam	Female	Age	STEM
Simple condition					
Accuracy	− 0.02 (0.06)	0.00 (0.05)	0.02 (0.05)	− 0.01 (0.00)**	0.03 (0.05)
Profit	− 0.01 (0.07)	0.00 (0.05)	− 0.01 (0.05)	− 0.01 (0.00)**	− 0.01 (0.06)
Complex condition					
Accuracy	− 0.02 (0.04)	− 0.03 (0.04)	− 0.01 (0.04)	− 0.00 (0.00)	0.08 (0.03)**
Profit	− 0.13 (0.10)	− 0.16 (0.10)	0.08 (0.09)	0.01 (0.01)	0.27 (0.08)***

"Accuracy" stands for the "Decision Accuracy Rate" metric, and "Profit" stands for the "Normalized Profit Ratio" metric. Data from all three formats is used. Standard errors clustered at the subject level in parentheses. Significance is denoted: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table 5 The three diagnostic tasks

Task	Experiment format		
	No Webcam	Webcam	Lab
Risk preference	5.06 (0.18)	5.13 (0.20)	5.19 (0.26)
CRT	1.62 (0.14)	1.68 (0.13)	1.91 (0.15)
Hit-15	0.83 (0.09)	0.99 (0.09)	0.89 (0.11)

Standard errors are in parentheses. The number in the Risk Preference task is the # of the lottery that serves as the "turning point" where subjects switch from preferring the lottery to the fixed amount. The number in the CRT task is the number of correct answers to the three questions in this task. The number in the Hit-15 task is the number of correct answers to the two questions designed to test subjects' understanding of the winning strategy in this task

comparisons).[15] However, we note that as previously, subjects in the No Webcam protocol have directionally the worst performance in the two tasks with objectively "correct" solutions (CRT and Hit-15).

Table 6 reports the results of regressing diagnostic task performance on treatment dummies and demographic variables (as in Function 1). We find that the two online protocols lead to comparable performance as the offline (lab) experiment for the risk preference task and the Hit-15 task. For the CRT, the No Webcam protocol leads to worse performance with weak significance ($p = 0.09$). Hence, in line with our previous results, we conclude that both online experimental protocols are capable of

[15] In the risk preference task, for all three formats more than 94% of subjects have a single switching point. The analysis here includes only subjects with a single switching point. If instead we consider all subjects and count the number of safe choices, all the results here hold. There is no significant pairwise comparison among the three experiment formats, nor are any estimates significant in the analogous regression to Table 6.

Table 6 Diagnostic tasks regression analysis

Task	Estimates for the independent variables				
	No Webcam	Webcam	Female	Age	STEM
Risk	− 0.17 (0.33)	− 0.06 (0.32)	− 0.42 (0.28)	0.03 (0.04)	0.08 (0.24)
CRT	− 0.33 (0.19)*	− 0.30 (0.18)	− 0.33 (0.16)**	− 0.04 (0.01)***	0.71 (0.15)***
Hit-15	− 0.03 (0.14)	0.10 (0.14)	0.11 (0.12)	− 0.01 (0.01)	0.03 (0.11)

Standard errors clustered at the subject level in parentheses. Data from all three formats are used. Significance is denoted: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 7 Dynamic resource allocation decision coefficient of variation analysis

Coefficient of variation	Experiment format		
	No Webcam	Webcam	Lab
Simple condition	0.33	0.23	0.20
Complex condition	1.77	1.55	1.11

generating comparable results to the offline (lab) format (with the Webcam protocol slightly preferred).

3.4 Implementation recommendations: sample size comparison

When choosing between the two online experiment protocols, a natural question is how the differences between protocols affect the statistical power, and hence the required sample size, of the experimental design. To provide guidance on that decision, in this section, we use our experimental data to conduct a series of bootstrap simulations to estimate the statistical power of each protocol to detect a treatment difference between simulated simple and complex conditions. We focus on a subject's profit as the outcome variable, as we observed a large and significant profit difference between the simple and complex conditions for all three formats.

Our data suggest two factors that could affect the statistical power of an experiment similar to ours: potential differences in the *magnitude* of a treatment effect, and potential differences in the *variability* (and thus possibly the *noise*) in the data. For the first factor, Table 3 shows that the treatment effect (reduction in average profit in the complex treatment) is directionally larger in the online protocols compared to the offline (lab) protocol. For the second factor, the variability of subjects' profits is directionally larger in both online protocols - but especially the No Webcam protocol. This can be seen more directly in Table 7, where we present the coefficient of variation (CV) for the normalized profit ratio. Profit in the No Webcam protocol has the greatest relative variability in both treatments.[16]

[16] We use the Feltz and Miller test (Feltz and Miller 1996) to compare the CV of profit between protocols. We find that the No Webcam protocol has a significantly larger CV compared to the offline (lab) experiment in the simple condition (p value = 0.02) and directionally larger CV in the complex condition (p value = 0.28). None of the other pair-wise comparisons are significant.

We now describe the bootstrap simulation analysis we use to formally compare the statistical power of each protocol. We use the following general steps to conduct the bootstrap analysis, as in Kleinman and Huang (2016) and Peng et al. (2005).

1. Conduct independent resampling from the two conditions (simple and complex) for the given sample size under consideration.[17]
2. Perform statistical tests to compare the two bootstrapped samples; to be consistent with our original method in identifying the treatment effect, here we use the rank-sum test to compare the two bootstrapped samples.
3. Repeats steps 1 and 2 a thousand times for each considered sample size.[18]
4. The proportion of samples where the null hypothesis is rejected is the estimated statistical power.

Figure 1 shows the results of this bootstrap simulation using the data from the simple and complex treatments for each protocol. This exercise helps us examine the statistical power of a protocol to detect the size of the effect observed in that protocol. By this criterion, each protocol achieves a reasonable level of statistical power. All three protocols would require fewer than 30 subjects to achieve 80% power to detect the empirically observed effect size for that protocol at a 5% significance level. Comparing across protocols, the two online protocols have higher power compared to the offline (lab) experiment. This is largely driven by the online formats having a larger treatment effect size (we will confirm this through an additional simulation analysis discussed below). Comparing between the two online protocols, the Webcam protocol has higher statistical power than the No Webcam protocol, consistent with the smaller variability of subjects' profits in that format. In summary, this analysis suggests that both online formats can achieve comparable statistical power to the offline format (and hence will require comparable sample sizes), but that between the online formats the Webcam protocol is preferable as it has better power.

However, the above analysis is not quite an apples-to-apples comparison, as it only asks whether a protocol has good power to detect the effect size observed in its own data. Because the protocols led to different treatment effect magnitudes, a protocol's performance is driven by two factors: (1) the difference in treatment effect sizes; (2) the difference in the variability of subject profit. Our preliminary comparison suggested that the primary potential issue with the No Webcam protocol is the greater noise in the data. We, therefore, conduct a follow-up simulation analysis to isolate the effect of variability alone on statistical power by holding fixed the conjectured size of the treatment effect across protocols. By varying the conjectured

[17] For each sample size, we allocate $\frac{1}{3}$ of the sample size to the simple condition and $\frac{2}{3}$ to the complex condition to account for the greater variability in the complex condition. The "sample size" or "number of subjects" in the text and figures all represent this *total* number of subjects from both conditions.

[18] This number of samples per considered sample size is in line with past research. Kleinman and Huang (2016) used 100 simulated samples per sample size, while Peng et al. (2005) used 1000 simulated samples.

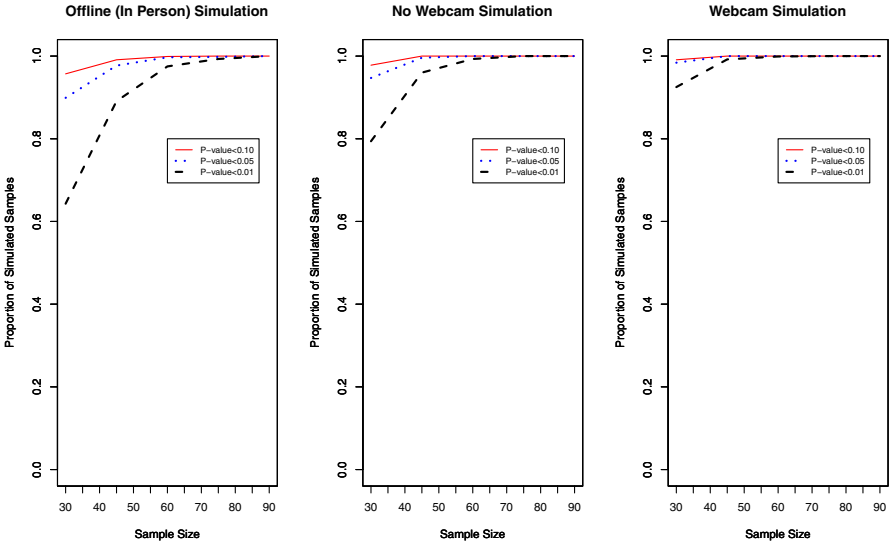

Fig. 1 Sample size simulation, empirically observed effect size

treatment effect size from small to large, we can examine how decision noise affects statistical power for each experimental protocol across a range of situations.

For this bootstrap simulation, we follow the general steps outlined above. However, before we conduct the resampling from the two conditions in Step 1, we first *demean* the two conditions such that they have a common mean of 0 for subjects' normalized profit ratios. Then, we add an adjustment factor $\delta > 0$ to subjects' profits in the simple condition to create a "conjectured treatment size". With this, we proceed to the resampling from the two (adjusted) conditions, and the rest of the bootstrapping procedures are the same. This approach allows us to equalize the (conjectured) treatment effect size across the three protocols and *solely* focus on the effect on power coming from the different variability of subjects' performance (i.e., normalized profit ratio) in each protocol. Note that this is similar to the typical approach taken when estimating the likely statistical power for a new treatment design with an unknown effect size. We consider δ ranging from in value 0.35 to 0.55 to capture small to large treatment effect sizes.

The simulation results are summarized in Figs. 2, 3 and 4. We observe that at $\delta = 0.35$, the No Webcam protocol is slightly weaker in power compared to the other two experiment formats although the differences across formats are not so salient: With around 50 subjects, all three formats can achieve a power of 80% at a 5% significance level. For $\delta = 0.45$, the No Webcam protocol has noticeably weaker power compared to the other two protocols. Specifically, to achieve 80% power to detect this moderate effect size, the No Webcam protocol requires more than 30 subjects while the other two protocols both require fewer than 30 subjects. Finally, for large effect sizes ($\delta = 0.55$), all three formats require fewer than 30 subjects to achieve a power of 80%, but the No Webcam protocol continues to have less power compared to the other two formats.

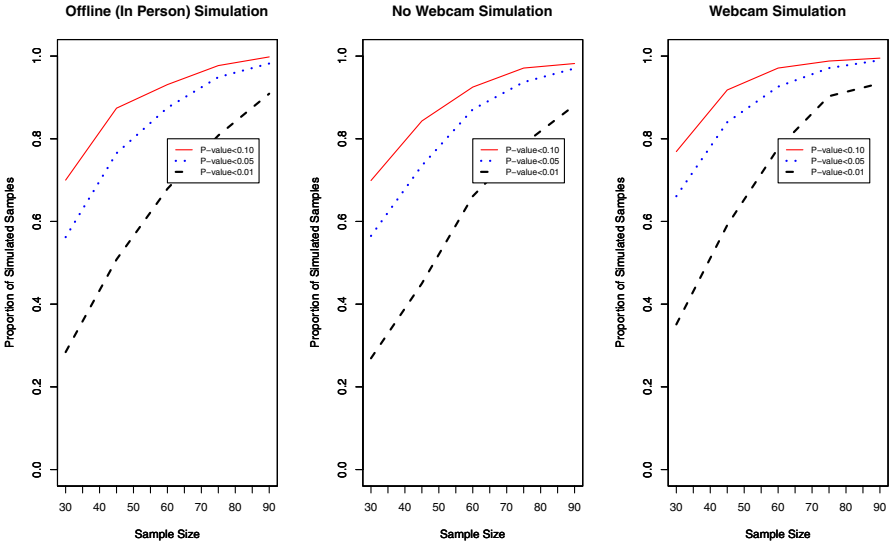

Fig. 2 Sample size simulation, treatment effect adjustment factor $\delta = 0.35$

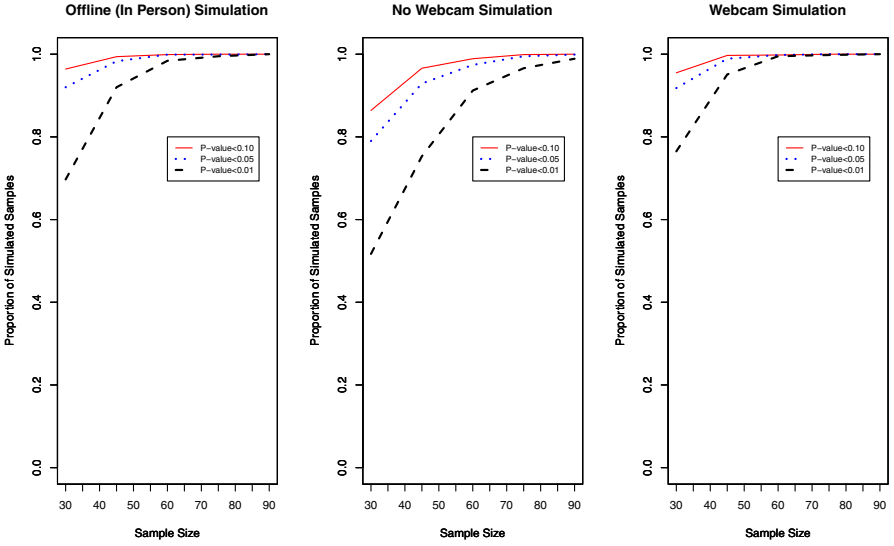

Fig. 3 Sample size simulation, treatment effect adjustment factor $\delta = 0.45$

In summary, both online protocols can give acceptable statistical power. However, the results here further highlight the consequence of the larger variation of subject performance in the No Webcam protocol and suggests that, between the two online protocols, the Webcam protocol is again the preferable format—with performance comparable to the offline (lab) experiment.

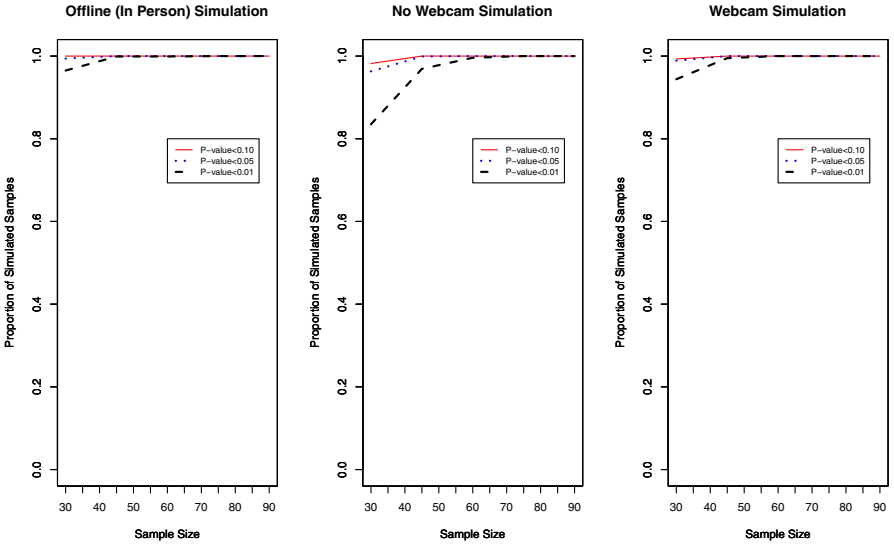

Fig. 4 Sample size simulation, treatment effect adjustment factor $\delta = 0.55$

4 Discussion and conclusion

In this paper, we discuss a novel protocol to combine an experimental platform like ZTREE Unleashed with web-conferencing software like Zoom to conduct the kind of complicated experiments typically done in an offline laboratory format. We show that both methods provide comparable data to the lab setting, both in terms of subject demographics and performance; however, we also show that the format with subjects' webcams on yields somewhat less noisy data (yielding better statistical power) and hence should be preferred. We contribute to the best practices in conducting online experiments both in terms of experiment protocols (webcam on versus off) as well as estimating statistical power and the selection of sample size. Our research is particularly relevant for researchers who wish to continue their research agendas online using ZTREE with the same subject pool.

We note that our four decision tasks are all individual decision-making tasks. For experiments with interactions among subjects, several of our results provide very positive indicators that such experiments can be successfully run online as an alternative or complement to in-person lab studies: Subjects are able to keep focused in relatively long and complicated economic experiments, and subjects dropping out is not an issue in our study. Hence, we believe that our online experiment formats will be able to generate high quality data for experiments with interactions. However, in strategic settings, researchers' decision about requiring subjects to use their webcam is more complicated. In experiments with strategic interactions, the webcam may create a separate channel for subjects to communicate with each other or to infer play based on facial reactions, potentially resulting in undesirable consequences. This can be partially mitigated by having subjects in separate virtual breakout rooms by role or fully mitigated by having individual breakout rooms. However, this

increases experimenter overhead for monitoring, as, for example, when using Zoom they will need a separate Zoom instance for each breakout room. If this is deemed infeasible given the researcher's context, our results suggest that utilizing the No Webcam protocol is also feasible, although potentially with a larger sample size.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40881-021-00112-w.

Funding The researchers were supported by the faculty research grant at the University of Michigan.

Availability of data and material (data transparency) All the experiment data, except for the subject information that needs to be anonymized, is available upon request from the authors.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability (software application or custom code) The experiment program files and codes are available from the authors upon request.

References

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*(1), 99–131.

Arechar, A. A. & Rand, D. G. (2021). Turking in the time of covid. *Behavior Research Methods, 53*, 2591–2595.

Carpenter, J., Graham, M., & Wolf, J. (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior, 80*, 115–130.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature, 47*(2), 448–474.

Duch, M. L., Grossmann, M. R., & Lauer, T. (2020). z-Tree unleashed: A novel client-integrating architecture for conducting z-Tree experiments over the Internet. *Journal of Behavioral and Experimental Finance, 28*, 100400.

Feltz, C. J., & Miller, G. E. (1996). An asymptotic test for the equality of coefficients of variation from k populations. *Statistics in Medicine, 15*(6), 647–658.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Gneezy, U., Rustichini, A., & Vostroknutov, A. (2010). Experience and insight in the race game. *Journal of Economic Behavior & Organization, 75*(2), 144–155.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association, 1*(1), 114–125.

Hauser, D., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. *Handbook of Research Methods in Consumer Psychology*, pages 319–337.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*(5), 1644–1655.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399–425.

Howell, D. C. (2012). *Statistical Methods for Psychology*. Cengage Learning.

Kleinman, K. & Huang, S. S. (2016). Calculating power by bootstrap, with an application to cluster-randomized trials. *EGEMs*, 4(1).

Kovalchik, S., Camerer, C. F., Grether, D. M., Plott, C. R., & Allman, J. M. (2005). Aging and decision making: A comparison between neurologically healthy elderly and young individuals. *Journal of Economic Behavior & Organization, 58*(1), 79–94.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188.

Peng, X., Peng, G., & Gonzales, C. (2005). Power analysis and sample size estimation using bootstrap. *Phoenix: Paper presented at PharmaSUG*.

Zhao, S., Vargas, K., Friedman, D., & Chavez, M. (2020). UCSC LEEPS lab protocol for online economics experiments. *SSRN Electronic Journal*.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493.