Experimental Economics (2021) 24:2–30
https://doi.org/10.1007/s10683-020-09658-8

ORIGINAL PAPER

Best practices in replication: a case study of common information in coordination games

Roy Chen[1] · **Yan Chen**[2,3] · **Yohanes E. Riyanto**[4]

Received: 1 May 2019 / Revised: 20 April 2020 / Accepted: 29 April 2020 / Published online: 26 May 2020

Abstract

Recently, social science research replicability has received close examination, with discussions revolving around the degree of success in replicating experimental results. We lend insight to the replication discussion by examining the quality of replication studies. We examine how even a seemingly minor protocol deviation in the experimental process (Camerer et al. in Science 351(6280):143–1436, 2016. https://doi.org/10.1126/science.aaf0918), the removal of common information, can lead to a finding of "non-replication" of the results from the original study (Chen and Chen in Am Econ Rev 101(6):2562–2589, 2011). Our analysis of the data from the original study, its replication, and a series of new experiments shows that, *with common information*, we obtain the original result in Chen and Chen (2011), whereas *without common information*, we obtain the null result in Camerer et al. (2016). Together, we use our findings to propose a set of procedure recommendations to increase the quality of replications of laboratory experiments in the social sciences.

Keywords Replication · Common information · Coordination games · Group identity

JEL Classification C71 · C91 · D71

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10683-020-09658-8) contains supplementary material, which is available to authorized users.

✉ Yan Chen
yanchen@umich.edu

[1] Department of Economics, National University of Singapore, Singapore 117570, Singapore

[2] School of Information, University of Michigan, Ann Arbor, MI 48109-2112, USA

[3] Department of Economics, School of Economics and Management, Tsinghua University, Beijing 100084, China

[4] Division of Economics, Nanyang Technological University, Singapore 639818, Singapore

1 Introduction

Over the past decade, the issue of empirical replicability in the social sciences has received a great deal of attention (Open Science Collaboration 2015; Camerer et al. 2016). In particular, this discussion has focused on the degree of success in replicating laboratory experiments, the interpretation when a study fails to be replicated (Gilbert et al. 2016), and the development of recommendations on how to approach a replication study (Coffman et al. 2017; Shrout and Rodgers 2018). This paper contributes to our collective understanding of the best practices for replication studies.

Among the large-scale replication studies, Camerer et al. (2016) examines the replicability of laboratory experiments in economics. They re-examined 18 experimental studies published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014, and successfully replicated 11 studies.

Of the seven experiments which did not replicate, Chen and Chen (2011) study whether an ingroup identity can help subjects overcome coordination failure and achieve efficient coordination in the minimum-effort game (Van Huyck et al. 1990). In the laboratory, subjects are randomly assigned to minimal (random, anonymous, non-interacting) groups. They then use an online chat program to solve a problem together with their group members. This online chat is designed to enhance their group identity. The authors establish both theoretically and experimentally that an enhanced group identity can lead to more efficient coordination in environments that would naturally lead to coordination failure (Goeree and Holt 2005). As this experimental result did not replicate, we investigate the causes of this specific case of non-replication and use this case to propose a set of replication procedures designed to increase the quality of replication efforts.

Examining potential reasons for this lack of replication, we first consider the possibility that cultural differences may play a role in the different results across the original and replication studies (Roth et al. 1991; Henrich et al. 2001). Specifically, the original study was conducted at the University of Michigan (UM), whereas the replication was conducted at the National University of Singapore (NUS), two universities with different ethnic and racial compositions (Online Appendix C). Indeed, the replication report notes that "Singapore subjects were less active in the communication stage than the American subjects" (Ho and Wu 2016).[1] To enhance ingroup identity in the study, subjects were asked to use a chat protocol to identify painters. Therefore, we began our investigation by comparing the effects of using (culturally appropriate) Chinese paintings versus the original study's German expressionist paintings (Chen and Li 2009) on both communication volume and subsequent effort among NUS students. However, interestingly, when the first author conducted baseline experiments at NUS using the UM protocol, the original results replicated [see Fig. 1 and ingroup coefficient = 27.24, $p < 0.001$, column (2) in Table 2].

We next examined the instructions provided to participants in the replication experiment. In the original study, subjects were given a paper copy of the full set of

[1] All replication reports are archived at http://experimentaleconreplications.com/replicationreports.html.

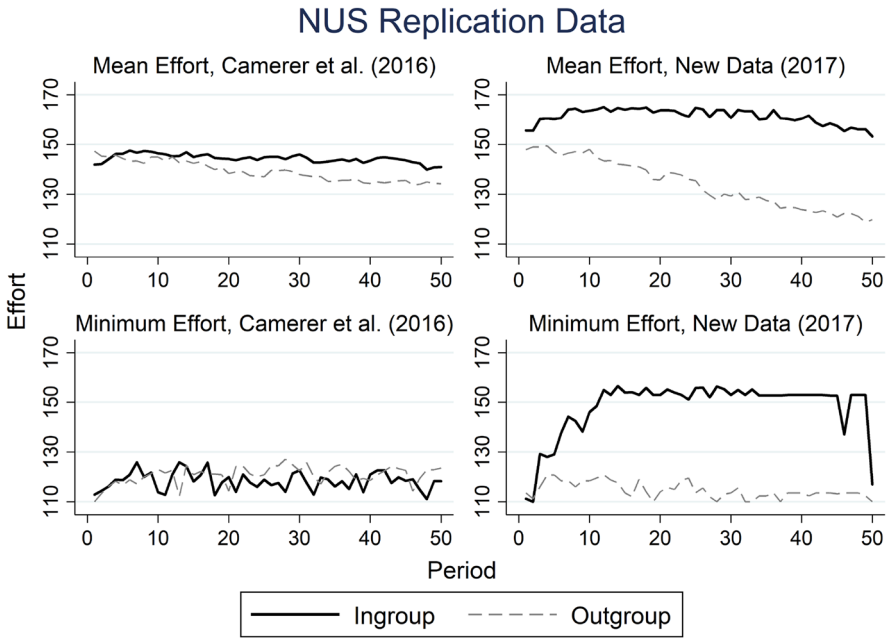

Fig. 1 The minimum and mean effort level across treatments at NUS

experimental instructions. However, this was not done in the replication. A recent experiment varies instruction delivery and reinforcement and shows that providing paper instructions is among the most effective methods to reduce potential subject misunderstandings during the experiment (Freeman et al. 2018).

Finally, we examined the provision of written versus oral instructions to participants. Subjects in the original experiment received a paper copy of the full set of instructions, were able to read the instructions on a computer screen, and had the instructions read aloud to them as well. We discovered after communications with the replication author that participants in the NUS replication study did not receive oral instructions for the minimum-effort game.[2] This is an important protocol deviation, since having subjects read these instructions on their individual computer screen without hearing the instructions aloud removed the common information condition of the original study. As it is impossible to implement common knowledge in the laboratory, we use *common information* to refer to the information condition where every subject knows the game form as well as what other subjects know about the game (Smith 1994). As such, it is an approximation of the *common knowledge* condition required by the game theoretic framework. Removing the oral instruction

[2] This protocol deviation can also be inferred from the Replication Report (http://experimentaleconreplications.com/replicationreports.html): "We will include one group of 12 subjects from each of the two treatments in each session" (Ho and Wu 2016). The two treatments have different instructions for the minimum-effort game, which cannot be read aloud simultaneously.

protocol without replacing it with another method to achieve common information means that subjects are no longer certain about what other subjects know, which increases the uncertainty surrounding others' strategies (Crawford 1995).

Previous research in experimental economics has outlined the importance of providing experimental instructions that match the information condition required by the respective theory. For experiments that require common information, experimenters read the instructions aloud to ensure that everyone in the lab is certain about what everyone else knows. Indeed, in an experimental study of advice-giving in coordination games, Chaudhuri et al. (2009) find that common information of advice leads to efficient coordination when advice from previous players is "not only distributed on paper but actually read out loud." They further find that even small deviations from the common information protocol lead to suboptimal outcomes. This leads to our second conjecture that common information is critical to establishing efficient coordination with ingroup members.

The rest of the paper is organized as follows. Section 2 reviews various experimental methods to approach common knowledge in the laboratory. In Sect. 3, we present our research design and summarize features of all four experiments used in our data analysis. In Sect. 4, we present our pre-registered hypotheses, our data analysis and results. Section 5 concludes by recommending a set of best practices for replication studies.

2 Approaching common knowledge in the laboratory

Common knowledge of the game form and rationality is assumed for various solution concepts in game theory (Geanakoplos 1992; Brandenburger 1992). Whether common knowledge exists in a given situation is important even in the most basic game theory analyses. A prominent example is that of the Centipede game (Rosenthal 1981). In a series of notes, Aumann (1995, 1996, 1998) and Binmore (1994, 1996, 1997) discuss whether the result obtained through backward induction (that the first players ends the game at the first opportunity) occurs as long as all players have common knowledge of rationality. Perhaps more relevant to the present paper, another example is the coordination game. In this game, players' try to coordinate their actions to achieve a mutually desirable outcome. Successful coordination depends crucially on players' mutual understanding about each others' behavior. The necessary condition for this is the presence of common knowledge regarding the nature of the game.

It is also clear that while common knowledge can affect fundamental game theory concepts in such a setting, it is difficult to achieve common knowledge in practice. One reason put forward by Morris and Shin (1997) is that people's belief about others' knowledge about the game are often derived from empirical sources, which are often inaccurate. In such a situation, they argue that common knowledge would be an excessively strong requirement that is impossible to achieve. This gives rise to imperfect coordination. In this regard, how can we approximate common knowledge to sustain a reasonable level of coordination? This question is highly relevant for any laboratory investigation of coordination games.

In what follows, we review the methods experimentalists have used to approach common knowledge in the laboratory.[3]

The most commonly used method is to read common instructions publicly out loud. Freitas et al. (2019) argue that common knowledge need not always require an infinite order of reasoning. Instead, people can often infer common knowledge using a variety of perceptual or conceptual cues. An example of such a cue is the presence of common information, where people are sure that others receive the same information. One way of ensuring this is by reading aloud the experimental instructions in front of all participants. This method has been advocated by the pioneers of experimental economics and adopted in the early laboratories in Arizona, Bonn, Caltech and Pittsburgh.[4] Smith (1994) writes that "experimentalists have attempted to implement the condition of 'common knowledge' by publicly announcing instructions, payoffs and other conditions in an experiment" and notes that this method achieves common information but is not sufficient to achieve common knowledge. Examples include Plott and Smith (1978) on the posted- and oral-bid trading institutions, Roth and Murnighan (1982) in bargaining games, Hennig-Schmidt et al. (2011) in health care payment system experiments, and numerous others.

The read-aloud method is sometimes adapted for specific constraints. For example, when interacting subjects are not in the same room (or the same city), Hennig-Schmidt et al. (2008) read the instructions out loud and specifically add the following sentence in their instructions, "The other group receives the same information as you do and knows that you also get this information." Thus, the experimenter publicly announces that every subject receives identical instructions.

In the replication context, a particularly interesting variation of the read-aloud method uses video instructions, which would potentially have the benefit of being able to use the exact same instructions with audio played out loud during one treatment and in headphones in another to vary the presence of common information. In addition to using video instructions with audio played out loud, Romero and Rosokha (2019) have an incentivized quiz and grouped students based on whether they have passed the incentivized quiz. In one treatment, after the quiz, subjects received a screen that said "There are 10 subjects in your group. Every subject in your group correctly answered all the questions on the quiz." We view this as a step further towards achieving common knowledge.

Another method to approach common knowledge has the subjects read their instructions in private first. Once everyone is finished, the experimenter then reads aloud a summary of the key points in public to create common information (Fehr et al. 2013). It is interesting to note that Fehr et al. (2013) was among the 11 experiments which successfully replicated in Camerer et al. (2016). In this case, however, the protocol for creating common information was preserved in the replication:

[3] We posted our question on the ESA Discussion Forum, soliciting methods experimentalists use to approach common knowledge, and received many helpful comments and references. We thank our colleagues for their contributions. See Acknowledgements for details.

[4] Private communications with Heike Hennig-Schmidt, Charles Plott, Karim Sadrieh, Burkhard Schipper and Al Roth.

"subjects read the instructions—including control questions—followed by a verbal summary of the authority game given by the instructor. Since verbal summaries of the instructions are read aloud, treatment randomization within each of the sessions is not feasible, identical to the original study."[5] Thus, even within the same replication project, we observe considerable variation in the extent to which the exact information conditions are preserved, with one replication group preserving common information (Holzmeister et al. 2016b) and another not preserving it (Ho and Wu 2016).

Lastly, within the class of minimum-effort coordination games, common information has been shown to be a crucial condition for successful replication. For example, Weber (2006) demonstrates that, by starting with small groups that successfully coordinate, adding players who are aware of the group's history can create efficiently coordinated large groups. This represents the first experimental demonstration of large groups tacitly coordinated at high levels of efficiency. When "growing" larger groups, the author created common information of previous minimum effort by writing this information on the board. In a recent replication of this study, Yang et al. (2017) provided this information by displaying it on the subjects' private computer screen (page 666), which led to significantly lower effort level than that observed in Weber (2006).

In sum, when theory requires common knowledge, experimentalists use different methods to create common information, which reduces the strategic uncertainty faced by the subjects. For game theory experiments where the solution concepts require common knowledge, the common information condition should be preserved in the corresponding replication studies.

3 Research design: four experiments

We consider the entire sequence of studies to be valid, with variations on the presence/absence of common information, as well as the dominant culture. Therefore, our study consists of four experiments: the original UM study (Chen and Chen 2011), the NUS replication Camerer et al. (2016), a new NUS replication following the UM protocol, and a new NTU study. To test whether the form of the instructions impacts the results in our case study, our third author, who was not part of either the original or the replication study, conducted new laboratory experiments at the Nanyang Technological University (NTU), varying ingroup/outgroup, as well as common information/no common information. We then analyse the entire series of studies to increase statistical power. Features of each part, including the number of independent sessions, the number of subjects per session, information condition and study site are reported in Table 1. In what follows, we summarize the experimental protocol in each of the four studies.

[5] See the corresponding replication report (Holzmeister et al. 2016b) archived at http://experimentaleconreplications.com/replicationreports.html.

Table 1 Features of experimental sessions used in analysis

Institution	Published data		New data (2017)		
	UM (2011)	NUS (2016)	NUS	NTU	
Common information	Yes	No	Yes	No	Yes
Ingroup	3×12	7×12	5×12	7×12	7×12
Outgroup	3×12	7×12	5×12	7×12	7×12
No. of subjects	72	168	120	168	168

The 2017 NUS study has five independent sessions for each of the two treatments, as we encountered administrative issues to continue using the NUS lab to collect more data for this project

The experimental procedure in the original UM study is described in Section III of Chen and Chen (2011). The subjects' experimental instructions are archived on the website of the *American Economic Review* as part of the online appendix.[6] Our new NUS replication conducted in 2017 follows the same protocol as the original study.

The NUS replication protocol is documented in the replication report (Ho and Wu 2016), archived on the replication study website. However, the replication report does not mention the protocol deviation of (1) removing common information; or (2) not distributing paper copies of instructions.

3.1 Economic environment

In all four experiments, the payoff function, in tokens, for a subject i matched with another subject j is the following: $\pi_i(x_i, x_j) = \min\{x_i, x_j\} - 0.75 \cdot x_i$, where x_i and x_j denote the effort levels chosen by subjects i and j, respectively; each can be any number from 110 to 170, with a resolution of 0.01. Using potential maximization as an equilibrium selection criterion (Monderer and Shapley 1996), absent of group identities, we expect subjects to converge close to the lowest effort level, 110, which is reported by Goeree and Holt (2005).

Chen and Chen (2011) use a group-contingent social preference model, where an agent maximizes a weighted sum of her own and others' payoffs, with weighting dependent on the group categories of the other players. Therefore, player i's utility function is a convex combination of her own payoff and the other player's payoff,

$$u_i(x) = \alpha_i^g \cdot \pi_j + \left(1 - \alpha_i^g\right) \cdot \pi_i(x) = \min\{x_1, x_j\} - c \cdot \left[\alpha_i^g \cdot x_j + \left(1 - \alpha_i^g\right) \cdot x_i\right], \tag{1}$$

where $\alpha_i^g \in [-1, 1]$ is player i's group-contingent other-regarding parameter, and $g \in \{I, O, N\}$ is the indicator for the other player's group membership, which can be from an ingroup (I), outgroup (O) or control condition (N). Based on estimations

[6] The URL for the online appendix is https://assets.aeaweb.org/assets/production/articles-attachments/aer/data/oct2011/20091062_app.pdf.

of α_i^g from Chen and Li (2009), we expect that $\alpha_i^I > \alpha_i^N > \alpha_i^O$, i.e., a player will put more weight on the payoff of an ingroup match, followed by a match in the control condition, then followed by a match from an outgroup.

With group-contingent social preferences, the potential function for this game becomes $P(x_i, x_j) = \min\{x_i, x_j\} - 0.75 \cdot [(1 - \alpha_i^g)x_i + (1 - \alpha_j^g)x_j]$. Chen and Chen (2011) show that, in the limit with no noise, this potential function is maximized at the most efficient equilibrium if $\alpha^g > \frac{1}{3}$, and at the least efficient equilibrium if $\alpha^g < \frac{1}{3}$ (Proposition 5). Using a stochastic potential to better approximate the noisy experimental data, Proposition 4 implies that, with sufficiently strong group identities, ingroup matching leads to a higher average equilibrium effort than either outgroup matching or control (non-group) matching. This forms the basis for our main hypothesis.

We registered our hypotheses and pre-analysis plan at the AEA RCT Registry, including five hypotheses, the power calculation, and an analysis plan (Chen et al. 2017).

For our power calculation, we used the results from Chen and Chen (2011). That study had 72 subjects (in the relevant treatments) and yielded a p value of 0.033. To obtain 90% power requires a sample size of 167 subjects. We could have instead used the results from Camerer et al. (2016), which had 168 subjects and a p value of 0.571. However, to obtain 90% power from using Camerer et al. (2016) would have required 3928 subjects, which was not feasible for this study. We also note that, by design, some replications will fail. If the power is 90%, then 10% of studies will not replicate by chance.

3.2 Experimental procedure

In our NTU study, we closely follow the original experimental procedure in Chen and Chen (2011) for our ingroup and outgroup with common information treatments. For our ingroup and outgroup without common information treatments, we modify the experimental procedure slightly. Specifically, we let participants read the second part of the instructions on the minimum-effort coordination game themselves on their computer screen. Across treatments, the first part of the instructions on the group assignment and the group enhancement task is identical to the one implemented in the original study.

In sum, for our NTU replication study, we implement a 2×2 between-subject factorial design that includes combinations of with- and without common information, as well as ingroup and outgroup variations. In our analysis, we pool the data from these four sources.

In our NTU replication study, we conduct 7 independent sessions per treatment, giving us a total of 28 sessions. In these sessions, we distribute hard copies of the instructions, read the first part of the experimental instructions out loud, and ask subjects to follow along by reading the copy of their written instructions silently. The third author and his research assistant (RA) each conducted 14 sessions. Subjects are randomly assigned to sessions, which are subsequently randomly assigned to the two experimenters.

Prior to conducting these 28 experiment sessions, we run a pilot session implementing the ingroup with common information treatment, conducted by the first author of the original study. This author demonstrates the entire sequence of the experiment protocol to the third author and his research assistant who both observe and take notes to maintain the consistency between the original study and our NTU replication study. Data from this pilot session is excluded from our data analysis.

Each of the main experimental sessions consists of 12 subjects. Each subject randomly draws an envelope containing an index card with either red or green color. There are equal numbers of red and green index cards. A subject ID is printed on this index card. Based on the color of their index card, subjects are assigned to either the Red or the Green group. Each of these groups has 6 members. They are then ushered to their respective PC client terminal.

In the identity enhancement stage, subjects are given five pairs of paintings, and each pair consists of a painting by Paul Klee and a painting by Wassily Kandinsky. Both are well-known expressionist artists. Subjects are given answer keys containing information about the artist who painted each of these paintings. They are given five minutes to review these paintings. Subsequently, subjects are shown two more paintings and are told that each of these paintings could have been painted by either Klee or Kandinsky, or by the same artist. Subjects are then given 10 min to submit their answers about the artist(s) who painted the two artworks. Within these 10 min, they are given a chance to communicate with others from the same group through an online communication protocol embedded into the z-tree program for the experiment. It is up to subjects to decide whether or not to actively use this communication channel; they are not required to submit the same answers as those given by other members. Subjects earn 350 tokens, which is equivalent to SGD 1, for each correct answer given. They are told the correct answers at the end of the experiment once the minimum-effort coordination game is completed. Note that this protocol is the same as that in the other three studies that comprise the pooled data set.

Once the identity enhancement stage is completed, subjects proceed to the minimum-effort coordination game. They play the game for 50 rounds. In the ingroup treatment, in every round, subjects are randomly matched with another subject from the same group, while in the outgroup treatment, subjects are randomly matched with another subject from the other group. In the treatments without common information, the instructions for the minimum-effort coordination game are not read aloud.

Finally, at the end of the experiment, subjects answer a post-experiment questionnaire containing questions on demographics, past giving behavior, strategies adopted during the experiment, group affiliation, and prior knowledge about the artists and their paintings. Summary statistics for the survey (at each experimental site) are presented in Online Appendix C.

In total, we have 336 subjects in our NTU replication study. They are undergraduate students at Nanyang Technological University (NTU) from different majors from the sciences, engineering, business and economics, the social sciences, and humanities. The average earnings per subject is around SGD 11 (including a SGD 5 fixed show-up fee for participation). The average duration of the experiment is one hour. Participants

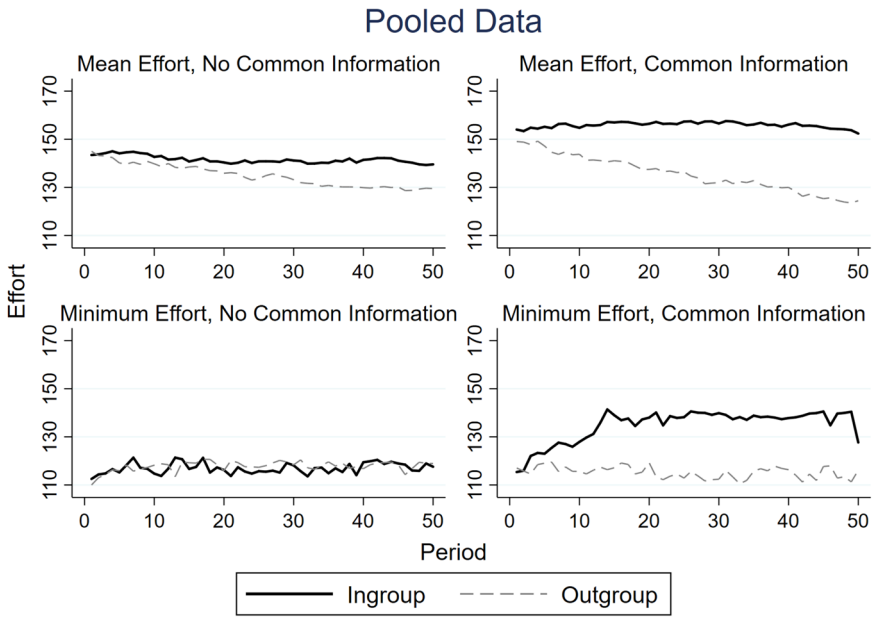

Fig. 2 The mean and minimum effort level across treatments, pooling date from all four studies

are allowed to participate in only one session and they remain completely anonymous throughout the experiment.

Our data analysis includes the original UM study with common information (Chen and Chen 2011), the NUS replication without common information (Camerer et al. 2016), the new NUS experiment with common information, and the new NTU study conducted both with and without common information. Overall, the data used in this study consists of 696 subjects in 58 sessions. Half of these sessions used ingroup matching while the other half used outgroup matching. Also, 30 of these sessions included common information while 28 did not include common information.

4 Results

In this section, we first present our pre-registered hypotheses and the corresponding analysis (Sect. 4.1). We then explore non-registered analyses on communication and learning dynamics (Sect. 4.2).

4.1 Pre-registered hypotheses and analysis

To examine whether common information plays a role in the results obtained, we first present our five pre-registered hypotheses and their corresponding results. Based on Proposition 4 in Chen and Chen (2011), we have the following hypothesis.

Hypothesis 1 With common information, average effort will be higher in the ingroup treatment than in the outgroup treatment.

Figure 2 presents the mean (top row) and minimum (bottom row) effort in the common information (right column) and no common information (left column) treatments. The ingroup (solid black line) and outgroup (dashed gray line) matching treatments are displayed together for each case. We see that the mean and minimum effort rises and remains stable over time only in the common information and ingroup treatment conditions.

Table 2 presents the results from a set of random-effects regressions separately for each study site (columns 1–3) as well as pooled (column 4), with standard errors adjusted for clustering at the session level. The dependent variable is the chosen effort level, while the independent variables include a dummy variable indicating whether a subject participated in an ingroup session. The top panel reports results under the common information condition. Column 1 shows the original result from Chen and Chen (2011) while column 2 of the bottom panel shows the replication result from Camerer et al. (2016).

Our pooled results in column 4 lead us to reject the null in favor of Hypothesis 1 (ingroup coefficient = 18.45, $p < 0.001$). The pooled result is consistent with the main result obtained in the original study (Chen and Chen 2011). It is also robust to adding demographic controls (Online Table B.1) or a time trend (Online Table B.2). Meanwhile, we note that results from the NTU study (column 3) are weaker. With common information, the ingroup dummy is marginally significant (ingroup coefficient = 9.85, $p < 0.10$), whereas it is not significant without common information. We also note that the magnitude of the coefficients are similar with and without common information. In the test of Hypothesis 5 towards the end of this subsection, we evaluate the heterogeneity of treatment effects across universities through their communication patterns, and we find that NTU subjects are the least communicative among the three subject pools.

Our next hypothesis is based on the protocol deviation in the replication study (Camerer et al. 2016).

Hypothesis 2 Without common information, average effort will be indistinguishable between the ingroup and outgroup treatments.

In the bottom panel of Table 2, our treatment dummy, ingroup, is not significant in any study location. Pooling all data (column 4), we fail to reject Hypothesis 2 (ingroup coefficient = 6.72, $p = 0.248$). This result affirms the results of the replication study (Camerer et al. 2016). Again, this result is robust to adding demographic controls (Online Table B.3) or a time trend (Online Table B.4)

Our next two hypotheses relate to the effect of common information when the matching protocol is kept constant.

Hypothesis 3 Between the ingroup treatments, average effort will be higher in the common information treatment than in the no common information treatment.

Table 2 Ingroup effects on effort at different universities and pooled: random-effects

Dependent variable: effort				
Common information (instructions displayed on individual terminals and read aloud)				
	(1) UM (2011)	(2) NUS (2017)	(3) NTU (2017)	(4) Pooled
Ingroup	23.85**	27.24***	9.85*	18.45***
	(11.200)	(6.391)	(5.561)	(4.658)
Constant	139.48***	134.11***	133.17***	134.75***
	(11.072)	(4.172)	(3.542)	(3.008)
Observations	3600	6000	8400	18,000
Number of subjects	72	120	168	360
R^2	0.2919	0.3449	0.0473	0.1553
No common information (instructions displayed on individual terminals but not read aloud)				
		(2) NUS (2016)	(3) NTU (2017)	(4) Pooled
Ingroup		5.20	8.23	6.72
		(9.197)	(6.927)	(5.819)
Constant		139.16***	130.54***	134.85***
		(7.459)	(5.386)	(4.664)
Observations		8400	8400	16,800
Number of subjects		168	168	336
R^2		0.0124	0.0332	0.0208

* ** and *** denote significance at the 10%, 5%, and 1% level, respectively

Table 3 presents the results of a series of random-effects regressions both separately (columns 1 and 2, NUS and NTU, respectively) and pooled (column 3), with standard errors adjusted for clustering at the session level. Note that we exclude the results obtained in the UM study from this set of regressions as all UM subjects received common information. The dependent variable for these regressions is again the effort level chosen, while the independent variables include a dummy variable indicating whether a subject participated in a session under common information.

From the pooled results in column 3 of the top panel, we see that the coefficient for common information is 11.63 ($p = 0.019$), leading us to reject the null in favor of Hypothesis 3. We note that this effect is primarily driven by NUS subjects (common information coefficient = 16.99, $p < 0.05$), whereas common information has no statistically significant effect on NTU subjects (common information coefficient = 4.25, $p > 0.10$).

Hypothesis 4 Between the outgroup treatments, average effort will be indistinguishable with or without common information.

Hypothesis 4 is based on potential maximization as an equilibrium selection principle in the minimum effort game (Monderer and Shapley 1996). With

Table 3 Common information effects in different universities and pooled

Dependent variable: effort			
Ingroup			
	(1) NUS	(2) NTU	(3) Pooled
Common information	16.99**	4.25	11.63**
	(7.235)	(6.111)	(4.976)
Constant	144.36***	138.77***	141.57***
	(5.416)	(4.355)	(3.478)
Observations	7200	8400	17,400
Number of subjects	144	168	348
R^2	0.1475	0.0078	0.0622
Outgroup			
	(1) NUS	(2) NTU	(3) Pooled
Common information	− 5.04	2.63	− 0.10
	(8.570)	(6.447)	(5.548)
Constant	139.16***	130.54***	134.85***
	(7.508)	(5.386)	(4.661)
Observations	7200	8400	17,400
Number of subjects	144	168	348
R^2	0.0118	0.0042	0.0000

Random effects regressions of common information on minimum-effort game choices

*, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. The top panel displays the sessions with ingroup matching and the bottom panel displays sessions with outgroup matching. The University of Michigan is not included because all sessions there included common information

For ingroup sessions at NUS, the common information coefficient is significantly different from 0 at the 5% level, while it is not at NTU. When these sessions are pooled, the coefficient is significant at the 5% level. For outgroup sessions, the common information coefficients are not significantly different from 0 at either university

group-contingent social preferences (Eq. 1), the minimum effort game continues to be a potential game, with its potential function maximized at the most (resp. least) efficient equilibrium if $\alpha^g > \frac{1}{3}$ (resp. $\alpha^g < \frac{1}{3}$) (Proposition 5 in Chen and Chen 2011). With outgroup matching, we expect α^g to be low with common information (Chen and Li 2009). Removing common information increases strategic uncertainty and therefore does not provide incentives to increase effort. Indeed our structural estimation of the stochastic fictitious play learning model yields $\alpha^g = 0.20$ with and $\alpha^g = 0.23$ without common information for outgroup matching (Table 5).

We present the results for our outgroup treatments in the bottom panel of Table 3. These results show that the coefficient for common information is not significant for the pooled data (common information coefficient $= -0.10$, $p = 0.985$). Therefore, we fail to reject Hypothesis 4 at the 5% level.

Overall, the above set of results suggests that ingroup matching can positively affect coordination only when common information is present. Furthermore, this finding can explain the difference between the results obtained in the original study (Chen and Chen 2011) and in the replication study (Camerer et al. 2016).

In addition to the main treatment effects, we also observe different communicativeness across institutions, reflected in our last hypothesis. Cultural differences in communicativeness in similar coordination games have also been noted (Brandts and Cooper 2007).

Hypothesis 5 With or without common information, Singapore subjects will be less active in the communication stage than American subjects.

On average, during the experiment, American (Singaporean) subjects each submitted 10.69 (6.62) lines to the chat. Treating each chat group of six subjects as one independent observation, a Mann–Whitney *U*-test shows that this difference is significant ($p < 0.01$). Using word counts instead of chat lines to test for communicativeness yields similar results (53.76 vs. 31.71 words, $p < 0.01$ one-sided rank-sum test). Therefore, we reject the null in favor of Hypothesis 5. Furthermore, within Singapore, we find that NUS subjects communicate more than NTU subjects (36.81 vs. 27.33 words, $p < 0.01$ one-sided rank-sum test). This may be related to the nationalities of the subject pools. According to the post-experimental survey, NUS consists of 71% Singaporeans, 11% Malaysians, and 18% others, while NTU consists of 62% Singaporeans, 22% Malaysians, and 16% others. The lack of communicativeness at NTU might explain the weakness of its ingroup treatment effect under common information (column 3 in Table 2).

4.2 Non-registered analysis: communication, learning, and experimenter effects

Based on our pre-registered analysis in Sect. 4.1, we will explore the effects of communication on learning dynamics, and experimenter effects. The analysis presented in this subsection is not pre-registered.

To understand how communicativeness affects subjects' effort provision in the coordination games, we decompose the effect of communicativeness into an effect on the initial conditions and an effect on learning dynamics during the game. Doing so, we first find that communication differences create different initial conditions in the minimum-effort game. In Table 4, we present two OLS specifications with first-period effort as the dependent variable, and individual chat characteristics as independent variables. Here, we find that the communication difference creates different initial conditions in the minimum-effort game. With ingroup matching (column 1), subjects who submitted more lines gave a significantly higher first period effort (coefficient for # of lines = 0.67, $p < 0.05$). This effect is not present with outgroup matching (column 2).

Secondly, we examine learning dynamics. In the theory of potential games (Monderer and Shapley 1996), of which the minimum-effort game is a special case, adaptive learning leads to convergence to the potential-maximizing

equilibrium (Blume 1993). We use a stochastic fictitious play learning model (Cheung and Friedman 1997) to uncover how treatments affect learning dynamics (Chen and Chen 2011). Table 5 presents the estimated parameters under different treatments. Here, we see that the combination of ingroup matching and common information increases subjects' group-contingent other-regarding preferences α (2-sided Mann–Whitney U-test, $p = 0.011$), while other determinants of learning dynamics, such as the discount factor (δ) and the sensitivity parameter (λ), are similar across sites.

Overall, our results here indicate that communicativeness impacts the initial-round effort in the coordination game but not learning dynamics, which instead are affected by a combination of ingroup matching and common information. These two effects together can explain the data from the original study (Chen and Chen 2011) as well as that obtained in the follow-up studies, including the replication (Camerer et al. 2016).

Finally, we examine any potential experimenter effects in the NTU study. In the UM and NUS studies, one experimenter ran all of the sessions in the study, whereas there were two experimenters at NTU: the third co-author, and his RA. Table 6 presents the results of a random-effects regression on the NTU data, with standard errors adjusted for clustering at the session level. The "RA" variable denotes sessions run by the RA.

This regression shows that, while the third co-author's sessions exhibit a significantly positive ingroup effect (ingroup coefficient = 19.10, $p < 0.05$), the RA's sessions show no such effect (RA + ingroup × RA = − 4.37, $p = 0.63$, Wald test). In addition, the effort levels of subjects in the outgroup treatment are marginally higher for the RA than for the third co-author. Comparing the outgroup efforts across studies (Table 2), the third co-author's sessions exhibit the lowest mean outgroup effort (125.36), partially accounting for the significant ingroup effect in these sessions. This analysis indicates that it is important to analyze experimenter effect when multiple experimenters conduct sessions.

5 Best practices in replication

Using the results from our analysis of the effect of common information in coordination games as well as reflections from the recent literature on replication from psychology and economics, we end by outlining a set of best practices for replicating economic, and perhaps other social science, experiments.

As a result of Open Science Collaboration (2015) and the resulting replication crisis, the field of experimental psychology has begun to confront several of the issues that we discuss in this article. The subsequent work by many psychologists closely relates to the situation that experimental economics is currently facing. Therefore, we incorporate some of the suggestions from the replication literature in psychology, e.g., Brandt et al. (2014), into the set of best practices for replication in experimental economics.

Table 4 Chat volume and period 1 effort

Dependent variable: effort in period 1		
	(1) Ingroup	(2) Outgroup
# of lines in chat	0.67**	0.27
	(0.311)	(0.221)
Painting 6 correct	2.80	3.54
	(2.212)	(2.582)
Painting 7 correct	− 0.75	1.51
	(3.551)	(4.076)
Analysis	− 0.72	− 0.79
	(0.581)	(0.663)
Question	− 0.95	0.12
	(1.297)	(1.006)
Agreement	− 0.41	− 0.35
	(1.269)	(1.032)
Engagement	0.68	1.34
	(1.591)	(1.640)
Constant	143.63***	139.73***
	(4.804)	(4.982)
Observations	348	348
R^2	0.025	0.013

OLS regressions of chat characteristics on minimum-effort game choices in the first period

*, **, and *** denote significance at the 10%, 5%, and 1% level, respectively

The "# of lines in chat" variable is a count of the number of lines submitted in the chat by each subject. "Painting 6 correct" and "painting 7 correct" are dummy variables regarding whether the subjects guessed the artists for the two additional paintings correctly. The variables "analysis" to "agreement" are a count of the number of lines by each subject that were categorized into each of the respective categories by trained chat coders. "Engagement" is a measure of how engaged each subject was in the chat, as coded by the chat coders. The number of lines submitted in the chat is positively and significantly (at the 5% level) correlated with first-period effort in the minimum-effort game, which immediately followed the chat, but only when there was ingroup matching. No other coefficient is significantly correlated with first-period effort

5.1 Exact versus close replication

We first differentiate between exact and close replications. An *exact replication* uses the same experimental protocol as the original study, including the same instructions, the same delivery of instructions and the same information conditions, but changes the experimenter and the subject pool.[7] The primary goal of an exact replication is to check whether an original finding is true, and can be observed by other experimenters. In other words, exact replications can be used by those in the scientific community to check the work of others, and potentially correct what we think

[7] Czibor et al. (2019) and Hamermesh (2007) call these "statistical replications."

Table 5 Learning model parameter estimates

	Discount factor	Sensitivity parameter	α (group-contingent, other-regarding parameter)			
			Common information		No common information	
	δ	λ	Ingroup	Outgroup	Ingroup	Outgroup
UM	0.70	3.36	0.73	0.40		
NUS	0.70	4.32	0.64	0.13	0.37	0.29
NTU	0.63	3.78	0.39	0.16	0.27	0.17
Overall			0.54	0.20	0.32	0.23

The δ and λ columns show the respective estimates for those parameters at the different universities separately. These estimates are quite similar to each other, suggesting that the learning dynamics do not change between experiment sites. The α estimates are generated for each session separately and the averages across sessions are displayed inthis table, separated by university and treatment

Table 6 Experimenter effects at NTU

Dependent variable: effort	
	NTU
Ingroup	19.10**
	(8.462)
Common information	5.59
	(4.741)
Ingroup × common information	− 1.43
	(12.617)
RA	14.65*
	(7.967)
Ingroup × RA	-19.02
	(12.120)
Common information × RA	− 2.02
	(9.133)
Ingroup × common information × RA	2.19
	(15.789)
Constant	122.17***
	(2.004)
Observations	16,800
Number of subjects	336
R^2	0.0947

Random-effects regression on the NTU data, with standard errors adjusted for clustering at the session level

*, **, and *** denote significance at the 10%, 5%, and 1% level, respectively. The "RA" dummy variable denotes when a session was run by the research assistant

we know. Exact replications are an ideal, and we argue that they can be achieved by replicators. For example, the replication of Fehr et al. (2013) by Holzmeister et al.

(2016b) is an exact replication, as the replicators kept the same instruction as well as the same delivery method of instructions, including its language (German) and the size of each replication session.

By contrast, a *close replication* must have protocol differences from the original study and document the differences explicitly. These are follow-up studies which can tell us how general and robust the original findings are.[8] By exploring the conditions under which the original results do or do not hold, we can understand the generalizability of a finding, or the bounds of its effectiveness. Close replications act as robustness checks on original studies.

Both exact and close replications provide valuable contributions to the social sciences, with advocates for both types of studies in psychology, economics, and business (e.g. Lynch Jr. et al. 2015; Coffman and Niederle 2015; Zwaan et al. 2018). The important point is that their difference, both in purpose and in effect, must be acknowledged. In particular, due to the potential of various "replicator degrees of freedom" (Bryan et al. 2019), or choice that replicators can make to create false negatives, close replications must not be incorrectly deemed to be exact replications.[9]

Clemens (2017) explores this difference, proposing that experimenters classify follow-up studies based on whether they estimate parameters drawn from the same (exact replication) or different (close replication/robustness) sampling distributions compared to the original study. Clemens (2017) also proposes that follow-up studies should be considered as robustness checks until the follow-up researchers prove that they are exact replications, since the "costs of spurious 'failed replications' for researchers are very high."

The stated goal of large-scale replication studies, such as the Open Science Collaboration (2015) and Camerer et al. (2016), is to be exact replications. These types of studies grab headlines precisely because they cause us to question findings that we previously took to be scientific truths. Protocol deviations such as the one identified in this paper change these studies into close replications. Their null findings are valuable in that they tell us that the deviated protocol strengthens group identity insufficiently to improve coordination, but they do not tell us whether the original results are true.

In what follows, we suggest best practices that enable exact replications, while acknowledging that close replications are valuable as well but are beyond the scope of this paper.

5.2 Best practices for replicators

When planning for an exact replication of an original study, the replicators must make the new study convincing. In particular, they must demonstrate that they have produced a faithful recreation of the original study with high statistical power.

[8] Czibor et al. (2019) and Hamermesh (2007) call these "scientific replications."

[9] Based on simulation results, the Replication Network published a dispute regarding whether Christoper Bryan p-hacked his original results: https://replicationindex.com/2019/12/02/christopher-j-bryan-claims-replicators-p-hack-to-get-non-significant-results-i-claim-he-p-hacked-his-original-results/.

Different replicators may have very different notions of how to achieve this, which can lead to varying quality in replication studies.

Inspired by the "Replication Recipe" in psychology (Brandt et al. 2014), we propose a set of questions replicators are expected to answer, and procedures that replicators should follow, and provide readers with a simple method of detecting differences between the replication and original studies. The questions highlight the common issues with replications that have been brought up in the past, and encourage the replicators to carefully consider these issues when designing their replications.

5.2.1 Understanding theory

Replicators should understand the theory underlying the experiment they plan to replicate.[10] Many experimental design details reflect the original experimenters' efforts to satisfy theoretical assumptions. Section 2 documents how experimentalists approach the common knowledge assumption underlying several game theoretic solution concepts. In coordination games, for example, abandoning the read-aloud method in favor of individual silent reading of instructions compromises the common information condition designed to approximate the common knowledge assumption, and upon which the solution concept (Nash equilibrium) rests. Understanding theory reduces the likelihood of protocol deviations in replication studies that compromise the internal validity of an experimental design.

Summarizing this subsection, we end with questions replicators should ask themselves when planning an exact replication:

Q1.1 What is the effect I am trying to replicate?
Q1.2 What are the assumptions in the original study that lead to the relevant hypotheses and the corresponding effect?
Q1.3 Will these assumptions also hold in my replication?

5.2.2 Choosing culturally appropriate subject pools

Though our study did not test this, there is evidence that culture matters in economic (and psychology) experiments. Henrich et al. (2001) show that across 15 small-scale societies, behavior in the ultimatum, public goods, and dictator games showed significant differences. This is likely context and game-specific. Roth et al. (1991) show in a four-country study that, while market experiments showed similar results everywhere (and matched theory), bargaining outcomes showed significant differences between countries (and none matched theory). Bavel et al. (2016) find that, among the studies that Open Science Collaboration (2015) examined, replication success was negatively correlated with the contextual sensitivity of the research topic.

For Chen and Chen (2011), intra-group communication and puzzle solving was designed to strengthen group identity sufficiently to improve coordination. The effectiveness of this intervention is likely affected by the cultures of the replication

[10] This point is less relevant for fact-finding experiments, which do not rely on theory.

site. As we see in our sessions, subjects in Singapore communicate with each other much less than those in the United States. This could partially be due to both the language in which the experiments were conducted, and the origins of the paintings used in the puzzle-solving task. The use of paintings by Chinese or other Asian artists may have increased the cultural appropriateness of the task and caused the subjects to communicate more with each other.

Appropriateness of experimental instruments is also an important consideration, and may have affected another study which did not replicate in Camerer et al. (2016). Specifically, Ifcher and Zarghamee (2011) used a film clip of Robin Williams to induce happiness in their subjects. However, between the original study and the replication study, Robin Williams had died of suicide. Also, the original study was conducted in the US while the replication study was conducted in the UK, two countries with different senses of humor.

In the replication of Duffy and Puzzello (2014), the original instructions, in English, were translated into German, which might have introduced different connotations depending on how words are translated. One good practice might be to have translated instructions translated back into the original language and to then make comparisons.

In at least three out of the seven "failed" replications in Camerer et al. (2016), the replication subject pool might not have been culturally appropriate. A simple rule of thumb is to conduct an exact replication in the same country or a culturally similar country. We summarize our discussions with the following questions:

Q2.1 In what country/region was the original study conducted?
Q2.2 In what language was the original study conducted?
Q2.3 Where was the original study conducted? (e.g., lab, field, online)
Q2.4 Who were the participants in the original study (e.g., students, Mturk, representative)?
Q2.5 Is my sample culturally appropriate given the experimental instructions?

5.2.3 Calculating sample size with sufficient statistical power

In order to mitigate false negatives, replication studies must have high power (around 90%). Assuming the same effect size as the original study, this is controlled by increasing the sample size of the replication. However, several studies have examined the issue of statistical power in replication studies and found that due to various factors, the stated power of these studies is likely incorrect and too low.

Many of these power issues are caused by between-study variation. McShane and Böckenholt (2014) point out that effect sizes are affected by various issues that we take for granted or assume should not matter, such as the social context, timing of the experiment, subject pool, etc. These cause standard power calculations to be overly optimistic. Their solution is to offer more conservative power calculation formulae which take into account between-study variation by including effect-size heterogeneity as an input. Similarly, Perugini et al. (2014) propose adjusting power calculations using "safeguard power." These calculations use the *lower bound* of the

confidence interval as the estimate of the effect size to achieve a better likelihood of correctly identifying the population effect size.

Maxwell et al. (2015) suggest that low statistical power comes from the prevalence of single replication studies, and that replicators should perform multiple replication studies in order to achieve the necessary power. Along these lines, Simons et al. (2014) propose the adoption of "Registered Replication Reports." These reports would compile a set of studies from a variety of laboratories that all attempt to reproduce an original finding. For very important original results, the increased cost is offset by the meta-analysis that becomes possible with these reports, allowing them to authoritatively establish the size and reliability of an effect. We summarize our main points in the following questions:

Q3.1 What is the effect size I am trying to replicate?
Q3.2 What is the confidence interval of the original effect?
Q3.3 What is the sample size of the original study?
Q3.4 What is the sample size and power of my replication study given the original effect size?
Q3.5 Is my replication sufficiently powered?
Q3.6 Does my power calculation use the *lower bound* of the confidence interval?

5.2.4 Randomizing across treatments and bundling sessions

A fundamental aspect of an experimental study is random assignment of experiment subjects across treatments. Loewenstein (1999) critiques experimental economists for their failure to assign subjects randomly to treatments, leading to a threat to internal validity. We acknowledge that Chen and Chen (2011) is subject to this critique, whereas our NTU study preserves random assignment without sacrificing common information. In what follows, we discuss different forms of random assignment in the lab setting.

For individual choice experiments and fact-finding experiments where common knowledge is not a requirement for the relevant hypotheses, subjects in each experimental session should be randomly assigned across treatments and stay in the same room during the experiment with the same experimenter [e.g., Maréchal et al. (2017)]. We call this practice *within-session randomization*, which has the advantage of taking care of potential session effects (Fréchette 2012). When subjects in a session are in multiple treatments, instructions can be displayed on each subject's private screen and perhaps accompanied by the experimenter's voice over headphones. Common information would be difficult to establish in such settings.

By contrast, for game theory and market experiments that require common knowledge for the relevant hypotheses, within-session randomization might compromise the critical information conditions of the underlying theory or introduce additional confounds. To preserve the common information condition of the original study, our NTU study randomized subjects who signed up for the same time slot into two sessions, one in each treatment and administered by either the third author or his research assistant. In other words, subjects were randomly assigned to one of two experimenters, which enables us to preserve the read-aloud method of the

original experiment and control for the timing of the experiment. The compromise, however, is the potential experimenter effect, which one can control for by adding experimenter fixed effects in the data analysis.[11]

In what follows, we use two examples from Camerer et al. (2016) to illustrate when not to use within-session randomization in exact replications. Both examples are game theory experiments.

In the first example, when replicating Chen and Chen (2011) at NUS, Ho and Wu (2016) used *within-session randomization across treatments*, where two sessions from two different treatments were run simultaneously in the same room. As the two treatments have different instructions, the compromise was not to read the instructions aloud. This led to a loss of common information, a crucial assumption in the theoretical framework. In an email correspondence discussing the successful replication of Fehr et al. (2013), Holger Herz wrote that "we asked for our treatments to be done across sessions, precisely because within session randomization makes creating common knowledge (either through reading instructions aloud or a summary aloud) difficult or maybe even impossible, depending on the experiment. … I think the question of creating common knowledge in the lab is a very important one in the context of the recent push towards within session randomization. There clearly is a trade-off." The method of within-session randomization should not be used when it compromises the creation of common information.

In the second example, when replicating Duffy and Puzzello (2014), Holzmeister et al. (2016a) bundled four sessions into one large session ($n = 24$) and used *within-session randomization within a treatment*, whereas the original study was conducted one session ($n = 6$) at a time. As the original experiment involved indefinite horizons, some groups were paused to allow for other groups experiencing longer horizons to finish. This protocol deviation is documented as follows, "As the number of periods was determined by rolling a die in advance, the overall length of the experiment differed among groups. On average, each group played 32.81 periods, with the largest difference in the number of periods being 6 within one session. In order to avoid that groups with lower numbers of periods have to wait for the other groups to finish, the zTree programs were paused for some of the groups between some periods such that all four groups finished almost simultaneously" (Holzmeister et al. 2016a). It is not clear how subjects interpreted such pauses, especially given that they were taking places at different times for different groups. In addition to boredom, the pause may have affected subjects' beliefs about the probability of continuation. The method of within-session randomization should not be used when it compromises the original protocols and introduces confounds, such as altered beliefs.

Common to both types of within-session randomization is the necessary bundling of multiple sessions into one larger session, which might lead to a loss in subject attention, introducing more noise into the data. Another potential consequence is the spread of the experimenters' attention. Duffy and Puzzello (2014) explicitly mentioned that "After the instructions were read, subjects had to correctly answer a

[11] Fréchette (2012) discusses causes of session effects, as well as the properties and adequacy of various standard solutions.

number of quiz questions testing their comprehension of the environment in which they would be making decisions. After all subjects had correctly answered all quiz questions, the experiment commenced with subjects making decisions anonymously using networked computer workstation." That is, the answers of the quiz were checked individually. By contrast, Holzmeister et al. (2016a) did not report that all of the 24 subjects answered all of the quiz questions correctly. When an original study administers quizzes, the replication study should administer them, check answers and report the proportion of correct answers.

We summarize our discussions and formulate the following questions for replicators:

Q4.1 How do I plan to randomize sessions across treatments?
Q4.2 Does my randomization compromise theoretical assumptions?
Q4.3 What is the size of an original experimental session?
Q4.4 What is the size of my experimental session?
Q4.5 What differences between the original study and my study might be expected to influence the size and/or direction of the effect?
Q4.6 What steps do I plan to take to test whether the differences listed above will influence the outcome of my replication attempt?

5.2.5 Obtaining original experimenters' endorsement

To facilitate exact replication, the replicators should first approach the original experimenters with their proposed replication plan. If there is disagreement regarding these replication procedures, the original experimenters should be given a chance to suggest the modifications required to achieve an exact replication.

First, sending the original authors a *pre-replication plan* is necessary but not sufficient to guarantee an exact replication. For example, for all the replicated papers in Camerer et al. (2016), the original authors were contacted and sent a pre-replication plan in December 2014. However, the procedure section of the replication plan of Chen and Chen (2011) contains only one sentence, "We plan to follow the exact procedure of the original article" (Ho 2014), which does not contain details about the actual implementation. At the minimum, the replicators should answer the questions from Q1.1 to Q4.6.

Second, the replicators should run and record a *pilot session*, which uses the same recruiting and randomization procedures, as well as the same session size as the planned actual replication sessions. The original authors should observe the actual pilot session in person if feasible or watch the recorded session.

We acknowledge that such coordination would be costly for original authors. One potential method of compensation is to use the number of successful replications of one's own work as a metric of one's scientific achievement (Coffman et al. 2017). Such incentives would push original authors towards cooperating with replicators and providing them with enough information to ensure that replications are close to exact.

5.2.6 Evaluating replication

While replication studies in experimental economics use classical null hypothesis significance testing to evaluate replication results, we note that several studies have proposed that replications take a Bayesian approach. Verhagen and Wagenmakers (2014) explore various Bayesian tests and demonstrate how previous studies and replications alter established knowledge through the Bayesian approach. When addressing how the field should consider failed replications, Earp and Trafimow (2015) consider the Bayesian approach and how failed replications affect the confidence of original findings.

McShane and Gal (2016) propose "a more holistic and integrative view of evidence that includes consideration of prior and related evidence, the type of problem being evaluated, the quality of the data, the effect size, and other considerations." They warn that null hypothesis significance testing leads to dichotomous rather than continuous interpretations of evidence, and that approaches such as confidence intervals and Bayesian modeling might have the same problems.

5.3 Best practices for original experimenters and journals

First, we consider what the experimenters should report in an original paper to facilitate exact replications. As stated by Palfrey and Porter (1991), "the relevant criterion is that enough detail be provided to enable another researcher to replicate the results in a manner that the original author(s) would accept as being valid." There are three major parts of a laboratory experiment that need to be clarified for any potential replicators: recruitment, pre-experimental procedures, and experimental procedures.

Regarding recruitment, the original experimenters should specify who the subjects are, including the population the subjects are drawn from. Depending on the experimental task, they might want to discuss the cultural norms of the subject pool (Henrich et al. 2010). Furthermore, the sign-up procedure should also be mentioned, as different methods of signing subjects up for experimental sessions can lead to different levels of filtering of the subject pool.

For the pre-experimental procedures, the original experimenters should report the circumstances the recruited subjects experience, including whether there is any communication between the experimenters and the subjects after they are invited but before they arrive at the laboratory; whether there are potentials for communication among subjects before the experiment begins. The latter could be affected by the availability of a waiting room, and whether subjects know each other before the experiment. The latter information can be collected through a post-experiment survey and could potentially affect subjects' social preference towards others in the session.

For the experimental procedures, the experimenters should again report the circumstances experienced by the actual participants. Specifically, there should be detailed information on instruction delivery: when are the subjects given the

instructions, and how are they delivered to the subjects? As we point out in Sect. 2, how the instructions are presented to subjects can have important effects on subject behavior. The experimenter's lab logs, which document questions asked and answers provided during all experimental sessions, are also possible items that should be included along with the instructions used. Furthermore, how subject comprehension is measured or ensured (e.g. practice rounds, quiz, questions) are useful information to indicate the extent to which common knowledge is approximated.

The prevalent practice is for authors to upload subject's instructions to an open-access archive. *Experimenter's instructions* (Online Appendix A) explicitly mark out what the experimenter does and says at different points throughout the experiment that are not listed in the subject's instructions. A classic example of an experimenter's instructions is presented in Online Appendix B of McKelvey and Palfrey (1992). The publication of the experimenter's instruction would help to clarify the method of instruction delivery. Protocol deviations, such as the ones in Ho and Wu (2016), are less likely to happen if replication authors use the experimenter's instructions. Journals should require authors to upload experimenter's instructions in addition to subject's instructions.

Lastly, as we have already discussed in Sect. 5.2.5, original experimenters should coordinate with replicators to share software, experimenters' instructions, and other experimental materials. They should observe the pilot session in person or watch a recorded pilot session, and suggest modifications required to achieve an exact replication.

6 Concluding remarks

Recent large-scale replication studies have caught a lot of attention to the issue of the replicability of social science research. In an attempt to understand why an original study failed to replicate, we uncover important protocol deviations in some of the replications. Drawing from the replication literature in psychology and economics, we differentiate between exact and close replications, and advocate for a set of best practices for exact replications for replicators as well as for original authors and journals.

Acknowledgements We thank Ala Avoyan, Colin Camerer, Alain Cohn, David Cooper, Rachel Croson, John Duffy, Lorenz Goette, Nagore Iriberri, H. V. Jagadish, Nancy Kotzian, Aradhna Krishna, Wolfgang Lorenzon, Muriel Niederle, Scott Page, Daniela Puzzello, Tanya Rosenblat, Al Roth, Stephanie Wang, Roberto Weber, and seminar participants at Michigan and Stanford for helpful discussions and comments. Our summary of how experimentalists approach common knowledge in the laboratory in Sect. 2 benefits from discussions with Ian Krajbich, Dorothea Kübler, Ron Harstadt, Holger Herz, Heike Hennig-Schmidt, Charles Holt, Charles Plott, Julian Romero, Karim Sadrieh, Burkhard Schipper, Katya Sherstyuk and Walter Yuan. We also thank Ruike Zhang for her research assistance.

Funding The financial support from the National University of Singapore (Economics Department Grant) and Nanyang Technological University (MOE Tier 1 Grant—M4011490) is gratefully acknowledged. This research was approved by the National University of Singapore IRB (Ref. Code A-16-150E) and the Nanyang Technological University IRB (Ref. Code IRB-2017-08-052).

Availability of data, material, and code Our study is registered at the AEA RCT Registry (AEARCTR-0002406). All data and experimental materials are included in the Electronic Supplementary Materials. The data and analysis codes reported in this paper have been deposited in the open Inter-university Consortium for Political and Social Research (ICPSR) data repository (https://doi.org/10.3886/E119065V1).

Compliance with ethical standards

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

References

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, *8*(1), 6–19. https://doi.org/10.1016/S0899-8256(05)80015-6.

Aumann, R. J. (1996). Reply to binmore. *Games and Economic Behavior*, *17*(1), 138–146. https://doi.org/10.1006/game.1996.0099.

Aumann, R. J. (1998). On the centipede game. *Games and Economic Behavior*, *23*(1), 97–105. https://doi.org/10.1006/game.1997.0605.

Bavel, J. J. V., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *PNAS*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113.

Binmore, K. (1994). Rationality in the centipede. In R. Fagin & M. Kaufmann (Eds.), *Proceedings of the fifth conference (TARK 1994)* (pp. 150–159). Retrieved May 24, 2020, from http://www.sciencedirect.com/science/article/pii/B9781483214535500143.

Binmore, K. (1996). A note on backward induction. *Games and Economic Behavior*, *17*(1), 135–137. https://doi.org/10.1006/game.1996.0098.

Binmore, K. (1997). Rationality and backward induction. *Journal of Economic Methodology*, *4*(1), 23–41. https://doi.org/10.1080/13501789700000002.

Blume, L. E. (1993). The statistical mechanics of strategic interaction. *Games and Economic Behavior*, *5*(3), 387–424. https://doi.org/10.1006/game.1993.1023.

Brandenburger, A. (1992). Knowledge and equilibrium in games. *Journal of Economic Perspectives*, *6*(4), 83–101. https://doi.org/10.1257/jep.6.4.83.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F.J., Geller, J., et al. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005.

Brandts, J., & Cooper, D. J. (2007). It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association*, *5*(6), 1223–1268. https://doi.org/10.1162/JEEA.2007.5.6.1223.

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*(51), 25535–25545. https://doi.org/10.1073/pnas.1910951116

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 143–1436. https://doi.org/10.1126/science.aaf0918.

Chaudhuri, A., Schotter, A., & Sopher, B. (2009). Talking ourselves to efficiency: Coordination in intergenerational minimum effort games with private, almost common and common knowledge of advice. *The Economic Journal*, *119*(534), 91–122.
Chen, R., & Chen, Y. (2011). The potential of social identity for equilibrium selection. *The American Economic Review*, *101*(6), 2562–2589.
Chen, R., Chen, Y., & Riyanto, Y. E. (2017). *Common knowledge in coordination games*. https://www.socialscienceregistry.org/trials/2406/history/21080. AEA RCT Registry. September 01.
Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *The American Economic Review*, *99*(1), 431–457.
Cheung, Y. W., & Friedman, D. (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior*, *19*(1), 46–76.
Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, *31*(1), 326–342. https://doi.org/10.1111/joes.12139.
Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, *29*(3), 81–98. https://doi.org/10.1257/jep.29.3.81.
Coffman, L. C., Niederle, M., & Wilson, A. J. (2017). A proposal to organize and promote replications. *American Economic Review*, *107*(5), 41–45. https://doi.org/10.1257/aer.p20171122.
Crawford, V. P. (1995). Adaptive dynamics in coordination games. *Econometrica*, *63*(1), 103–143.
Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432. https://doi.org/10.1002/soej.12392.
Duffy, J., & Puzzello, D. (2014). Gift exchange versus monetary exchange: Theory and evidence. *American Economic Review*, *104*(6), 1735–76. https://doi.org/10.1257/aer.104.6.1735.
Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*(621), 1–11. https://doi.org/10.3389/fpsyg.2015.00621.
Fehr, E., Herz, H., & Wilkening, T. (2013). The lure of authority: Motivation and incentive effects of power. *American Economic Review*, *103*(4), 1325–59. https://doi.org/10.1257/aer.103.4.1325.
Fréchette, G. R. (2012). Session-effects in the laboratory. *Experimental Economics*, *15*(3), 485–498. https://doi.org/10.1007/s10683-011-9309-1.
Freeman, D. J., Kimbrough, E. O., Petersen, G. M., & Tong, H. T. (2018). Instructions. *Journal of the Economic Science Association*, *4*(2), 165–179. https://doi.org/10.1007/s40881-018-0059-0.
Freitas, J. D., Thomas, K., DeScioli, P., & Pinker, S. (2019). Common knowledge, coordination, and strategic mentalizing in human social life. *Proceedings of the National Academy of Sciences*, *116*(28), 13751–13758. https://doi.org/10.1073/pnas.1905518116.
Geanakoplos, J. (1992). Common knowledge. *Journal of Economic Perspectives*, *6*(4), 53–82. https://doi.org/10.1257/jep.6.4.53.
Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243.
Goeree, J. K., & Holt, C. A. (2005). An experimental study of costly coordination. *Games and Economic Behavior*, *51*(2), 349–364. https://doi.org/10.1016/j.geb.2004.08.006 **(special Issue in Honor of Richard D. McKelvey)**.
Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics*, *40*(3), 715–733. https://doi.org/10.1111/j.1365-2966.2007.00428.x.
Hennig-Schmidt, H., Li, Z. Y., & Yang, C. (2008). Why people reject advantageous offers—Non-monotonic strategies in ultimatum bargaining: Evaluating a video experiment run in PR china. *Journal of Economic Behavior & Organization*, *65*(2), 373–384. https://doi.org/10.1016/j.jebo.2005.10.003.
Hennig-Schmidt, H., Selten, R., & Wiesen, D. (2011). How payment systems affect physicians' provision behaviour—An experimental investigation. *Journal of Health Economics*, *30*(4), 637–646. https://doi.org/10.1016/j.jhealeco.2011.05.001.
Henrich, J., Boyd, R., Bowles, S., Camerer, C. F., Fehr, E., Gintis, H., et al. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.
Ho, T. H. (2014). Replication of Chen and Chen (American Economic Review, 2011) **(unpublished)**
Ho, T. H., & Wu, H. (2016). *Replication of the potential of social identity for equilibrium selection*. Retrieved May 24, 2020, from http://experimentaleconreplications.com/replicationreports.html

Holzmeister, F., Huber, J., Kirchler, M., & Razen, M. (2016a). *Replication of gift exchange versus monetary exchange: Theory and evidence*. Retrieved May 24, 2020, from http://experimentaleconreplications.com/replicationreports.html

Holzmeister, F., Huber, J., Kirchler, M., & Razen, M. (2016b). Replication of the lure of authority: Motivation and incentive effects of power, http://experimentaleconreplications.com/replicationreports.html

Ifcher, J., & Zarghamee, H. (2011). Happiness and time preference: The effect of positive affect in a random-assignment experiment. *American Economic Review*, *101*(7), 3109–3129. https://doi.org/10.1257/aer.101.7.3109.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, *109*(453), 25–34. https://doi.org/10.1111/1468-0297.00400.

Lynch, J. G, Jr., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing*, *32*(4), 333–342. https://doi.org/10.1016/j.ijresmar.2015.09.006.

Maréchal, M. A., Cohn, A., Ugazio, G., & Ruff, C. C. (2017). Increasing honesty in humans with noninvasive brain stimulation. *Proceedings of the National Academy of Sciences*, *114*(17), 4360–4364. https://doi.org/10.1073/pnas.1614912114.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, *60*(4), 803–836.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513.

McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, *62*(6), 1707–1718. https://doi.org/10.1287/mnsc.2015.2212.

Monderer, D., & Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, *14*, 124–143.

Morris, S., & Shin, H. S. (1997). Approximate common knowledge and co-ordination: Recent lessons from game theory. *Journal of Logic, language, and Information*, *6*, 171–190. https://doi.org/10.1023/A:1008270519000.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*,. https://doi.org/10.1126/science.aac4716.

Palfrey, T., & Porter, R. (1991). Guidelines for submission of manuscripts on experimental economics. *Econometrica*, *59*, 1197–1198.

Perugini, M., Gallucci, M., & Constantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519.

Plott, C. R., & Smith, V. L. (1978). An experimental examination of two exchange institutions. *The Review of Economic Studies*, *45*(1), 133–153.

Romero, J., & Rosokha, Y. (2019). The evolution of cooperation: The role of costly strategy adjustments. *American Economic Journal: Microeconomics*, *11*(1), 299–328.

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, *25*(1), 92–100.

Roth, A. E., & Murnighan, J. K. (1982). The role of information in bargaining: An experimental study. *Econometrica*, *50*(5), 1123–1142.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in jerusalem, ljubljana, pittsburgh, and tokyo: An experimental study. *The American Economic Review*, *81*(5), 1068–1095.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. https://doi.org/10.1177/1745691614543974.

Smith, V. L. (1994). Economics in the laboratory. *Journal of Economic Perspectives*, *8*(1), 113–131. https://doi.org/10.1257/jep.8.1.113.

Van Huyck, J. B., Battalio, R. C., & Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, *80*, 234–248.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731.

Weber, R. A. (2006). Managing growth to achieve efficient coordination in large groups. *Amer Econ Rev*, *96*(1), 114–126.

Yang, C. L., Xu, M. L., Meng, J., & Tang, F. F. (2017). Efficient large-size coordination via voluntary group formation: An experiment. *International Economic Review*, *58*(2), 651–668. https://doi.org/10.1111/iere.12230.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*(E120), 1–61. https://doi.org/10.1017/S0140525X17001972.