Experience, learning and the detection of deception[*]

Priyodorshi Banerjee [†] Sanmitra Ghosh[‡] Sanchaita Hazra[§]

November 2022

Abstract

Deceptive communication or behavior can inflict loss, making it important to be able to distinguish these from trustworthy ones. This article pursues the hypothesis that repeated exposure or experience can cause learning and hence better detection of deception. We investigate using data culled from events in a TV game show. Decision-makers in the show repeatedly faced situations where they had to correctly identify an individual from within a group all claiming to be that individual. Our sample showed evidence of learning, suggesting that experience can aid in assessing the credibility of cheap talk.

JEL Classifications: C89, D82, D83, J24

Keywords: deception detection; experience; learning

[*] We are grateful to the Eastern Regional Centre of the Indian Council for Social Science Research for financial support. We thank Ritwik Banerjee, Haimanti Bhattacharya and Saibal Kar for helpful comments. The authors are responsible for all errors.

[†] Economic Research Unit, Indian Statistical Institute, Kolkata, West Bengal, India; *banpriyo@isical.ac.in, banpriyo@gmail.com*

[‡] Department of Economics, Jadavpur University, Kolkata, West Bengal, India; *sanmitra.ghosh@jadavpuruniversity.in, sanmitraz@gmail.com.*

[§] Department of Economics, University of Utah, Salt Lake City, Utah, USA; *sanchaita.hazra@utah.edu, sanchaitahazra@gmail.com.*

1 Introduction

Concern surrounding deceptive communication such as false testimony or 'fake news' has increased in the recent past, with research suggesting that deceptive communication or manipulated information can adversely influence political (e.g., Rose 2017; Fujiwara et al. 2021), social (e.g., Kim et al. 2020) and economic outcomes (e.g., Institutional Investor 2019). A question of importance is if the ability to correctly assess credibility and detect false communication can be augmented through repeated exposure to environments with potential deception. In this article, we focus on situations where individuals may be repeatedly exposed, and ask whether such experience can enable learning and thereby yield improved assessment.

When the preferences of a sender and a receiver are opposed, economic theory traditionally proposes that strategic communication should be uninformative (Crawford and Sobel 1982), implying deception in communication cannot be systematically detected. However, extensive research on deception (Zuckerman et al. 1981; DePaulo et al. 2003; George et al. 2004; Hartwig and Bond 2011) shows that senders may involuntarily generate behavioral cues, i.e., 'signal truth' (Ockenfels and Selten 2000), which in principle can help discriminate between sincere and opportunistic communication.[1] For example, Wang et al (2010) found that the dilation of senders' pupils around the moment of communication was positively correlated with the magnitude of their deception, implying more information was getting revealed than could be inferred from equilibrium properties.

The presence of such valid indicators of deception suggests that individuals may be able to observe and exploit these cues to produce superior assessment. The literature however finds that humans are not always proficient in this domain, with performance in deception detection tasks often at levels consistent with random decision-making (Ockenfels and Selten 2000; Brosig 2002; Serra-Garcia and Gneezy 2021). A key question is thus whether there are factors which can help reduce detection error. We investigate if experience can

[1] Such cues could be generated because the act of deception imposes affective or cognitive stress.

act as such a factor, by affording the possibility of learning.

It is reasonable to expect learning in the domain of deception detection. Human capital models (Mincer 1974; Becker 1975) suggest that experience augments productivity through on-the-job learning. Relationships predicted by these models have received empirical support for employees as well as entrepreneurs (e.g., Maranto and Rodgers 1984; Taylor 2001). Additionally, laboratory research has shown, for a variety of tasks, that experience can cause learning (see e.g., Newell and Rosenbloom 1981 and Erev and Haruvy 2016). In the current context, decision-makers may observe cues potentially associated with deception, with experience enabling learning of patterns and improving inference.

Deception by nature is difficult to observe in the field. As data source, we identify a TV game show instead.[2] The show was first aired in the U.S. in the late 1950s, and has many *episodes*, each with multiple *sessions*. In any session, *judges* face a group of *challengers*, one of whom is the *central character*, the others being *impostors*. The task of each judge is to independently identify which challenger is the central character, with the impostors attempting to deceive the judges. A judge benefits if his/her identification is correct, with no explicit competition between judges, while all challengers benefit equally for every judge making an error. The set of judges is fixed for an episode, and a person can appear as a judge in multiple episodes. Since judges make many decisions over time, the show allows us to observe the effect of experience on detection error in a high-stakes, quasi-naturalistic environment. Details of the show and our data are in Section 2.

Section 3 presents preliminary findings. We find here that detection error on average is mostly lower than the random choice benchmark level. We further find an apparent trend, yielding that increased experience reduces average detection error.

Were analysis to yield a significant negative relationship between experience and error, a possible explanation would certainly be learning. The nature of our data would also

[2] Data from game shows have been used in economics at least since Gertner (1993). Darai and Grätz (2010), Belot et al. (2012) and Turmunkh et al. (2019) have earlier used such data for the analysis of deception detection.

render two alternative explanations admissible. One is predicated on deception by some challenger groups being intrinsically easier to detect than by others, by which we mean that large random samples of judges would generate lower detection error for the easier groups. If so, another possible explanation for a potential negative relationship would be that the chronological arrangement of challenger groups is such that later episodes, or later sessions within episodes, involve easier groups. The second is predicated on some judges being intrinsically more able at detection than others, by which we mean that decisions across large random samples of challenger groups would display lower error for more able judges. If so, a possible explanation would be that the chronological arrangement of judges is such that later episodes involve more able ones. This confound is of importance as judges on the show could be repeated across episodes yet did not appear in an equal number of episodes. This raises the concern that the number of episodes per judge is driven by a selection process favoring better-performing judges, who get to appear more often. An explanation for a potential negative relationship could then be that more experienced judges are the more innately able ones, as they are selected more often.

In Section 4, we analyze if the apparent negative relationship between experience and error is significant and find that it is. We particularly analyze if learning can emerge as an explanation for this relationship after accounting for potential bias in the selection of judges into episodes and features of the specific arrangement of challenger groups across sessions and episodes. We present findings from multiple analytical routes and conclude there is sufficient evidence for experience positively affecting the chance of success in deception detection by enabling learning.

The hypothesis that experience improves detection due to learning has received extensive attention in psychology, starting with Kraut and Poe (1980). Its importance has been recognized in economics as well (e.g., Wang et al. 2010, p. 1005). Prior research has compared performance in laboratory tasks of lay individuals to that of professionals such as judges or law enforcement officers. Distinction between junior and senior professionals, determined by length of active service, has also been made (DePaulo and Pfeifer 1986). A problem with this approach is that professionals diverge from lay individuals and each other on multiple dimensions, such as the degree or type of experience and training. Hence differences across these groups can not be uniquely attributed to experience. More strikingly, only a few studies (e.g., Ekman and O'Sullivan 1991; Vrij and Mann 2001) have found that experience matters, most reporting no difference between groups.

This apparent inability to benefit from experience could be because existing laboratory studies have typically not allowed high stakes, feedback, or repeated decisions (e.g., Vrij and Mann 2001; Elaad 2003).[3] The question of ecological validity has also been raised (e.g., Gokhmann et al. 2012), given the importance of contextual information in professional decision-making (e.g., Park et al. 2002; Belot and van de Ven 2017). Our approach is novel as it uses non-experimental data and studies a high-stakes, feedback environment with long series of decisions by individual judges without special training, who face different challengers on every occasion. Further, our setting contains elements of naturalism, as the task is to determine the identity of a human, with judges aware of some facts about the person.

A central theme in the literature on deception detection is whether there are factors able to improve detection accuracy. The commonly discussed factor is training (Zuckerman et al. 1984), although meta-analyses give mixed results with respect to efficacy and cost-effectiveness (e.g., George et al. 2004; Hauch et al. 2016). Other identified factors are individual versus group decision-making (Klein and Epley 2015) and personal biochemistry (Pfundmair et al. 2017). Our paper contributes to the literature by showing that experience may be a valid factor aiding detection. With respect to the training versus experience debate relevant to the design of hierarchies, and compensation and promotion policies for professional law enforcement, security, intelligence, and judicial services, as

[3] In Masip et al. (2018), a potential deceiver is interviewed twice in a laboratory set up by the same interviewer, the second time without forewarning. The interviewer in their setting is thus allowed two decisions. However, they do not allow feedback or salient incentives. Our setting additionally differs in that potential deceivers face no surprises and are interviewed only once.

well as other organizations where credibility of internal and external communication is vital, our findings suggest that experience may be of importance, and on-the-job learning may yield reliability.

This essay may also contribute to emerging debates on organization of media and social networks. For example, with respect to whether fundamental reform of these networks, including regulation, enforcement and public ownership, is necessary (e.g., Bak-Coleman et al. 2021) to combat the proliferation of fake news and manipulated information, our results indicate that decentralized individual or organizational responses may have hitherto unrecognized power. This suggests that awareness campaigns being pursued in many European countries, Australia, Canada, etc. may have merit, provided information and guidelines transmitted through such campaigns effectively proxy for the experience that could have been obtained otherwise. A further suggestion is that initiatives to create watchdog or observer networks to discern deception or validate information, such as the Digital Media Observatory of the European Commission, can benefit by drawing on civil society members at large in addition to certified experts, provided mechanisms to determine experience and accuracy are available.

Before proceeding, note that the judges and challenger groups on the show could have been samples pre-selected from outside pools of potential participants. We have no information on such outside pools. This limitation restricts the range of data available for examination, although it does not affect inferences conditional on samples observed. Some other main limitations of our paper are discussed in the concluding Section 5.

2 The game show, and data

We consider the first season of the U.S. TV game show *To Tell The Truth*, which aired on CBS over December 1956 through December 1959.[4] There were 145 episodes in the

[4] All episodes of this season have been made publicly available by CBS on its youtube channel. The playlist is at `https://www.youtube.com/playlist?list=PL39ftvD_GHaHhv8Qm_truGRLp61iJNFOg`.

season, usually one per week over the period, each lasting for about 30 minutes. Each episode had multiple sessions, for a total of 429.

Every session had a host, with the host fixed for an episode. Most episodes had the same person in the role of the host. There were always four judges, two males and two females, in all sessions and episodes. The set of judges was fixed for an episode, and the same person could appear as a judge in multiple episodes. In all, 56 individuals appeared in the season in the role of judge, 21 females, and 35 males. The judges were celebrities, usually from the entertainment sector. Most sessions had three individuals appearing as challengers, one being the central character and the other two impostors. The set of challengers usually consisted of individuals of the same gender, with mixed gender groups appearing occasionally. A person could appear as a challenger only once in the season out of all sessions and episodes.

A session began with the host publicly revealing some true information about the central character.[5] It was commonly known from the beginning that the central character's identity is not known to the judges. The task of each judge was to identify which challenger is in actuality the central character. All judges could publicly question the challengers to pursue this task. The central character could not give a false answer to any question, while the impostors could. After all questions had been asked and answers received, judges had to independently and simultaneously submit judgments. It is likely that a significant part of the viewership of the show were fans or potential fans of the celebrities who appeared as judges. This suggests judges may have had strong incentives to perform well in order to acquire, impress or retain followers.[6] The challengers as a group got $250, to be divided equally, for every judge making a mistake in identification.[7]

[5] Further details on the chronology of events within a session can be found in Appendix A.

[6] We assume that the objective of judges is to correctly identify the impostors. Given the judges were mostly entertainment professionals, they could have had other objectives, such as to provide a good show. We do not have extraneous information on judge objectives, but our assumption should suffice as long as these include correct identification.

[7] A challenger could thus earn up to $333. These are high stakes as the average starting monthly salary for a college graduate in 1960 was around $500 (National Association of Colleges and Employers 2012).

The description above pertains to what we call regular sessions. We call an episode regular if it consists of 3 regular sessions. Most sessions and episodes were regular. A session could be irregular for the following reasons (there were 32 such irregular sessions):

(I) The host differed from the usual (4 episodes, covering 11 sessions).

(II) The challengers were couples (4 sessions across 4 episodes). In some sessions, there were 3 couples (1 male and 1 female) as challengers. There were 2 central characters, from a single couple. The other couples were the impostors.

(III) 3 judges submitted a judgment (17 sessions across 17 episodes). There were sessions in which a single judge recused himself/herself due to prior acquaintance with a challenger (there was never any instance in which more than 1 judge recused).

An additional source of irregularity was session length. The first 6 episodes of the season contained 2 sessions, each lasting for about 15 minutes. All other episodes contained 3 sessions, each lasting for about 10 minutes.

Variables capturing these irregularities are used, along with other session- and episode-level variables, in our regression analysis, reported in Section 4. In any session, a judgment is either correct or incorrect. Our main dependent variable is judge decision, which is binary. The main independent variable, capturing the degree of experience, is the number of sessions, including the session from which the observation is being drawn, in which a judge has appeared at the time of decision.[8] Our paper aims to identify and explain any possible correlation between these two variables.

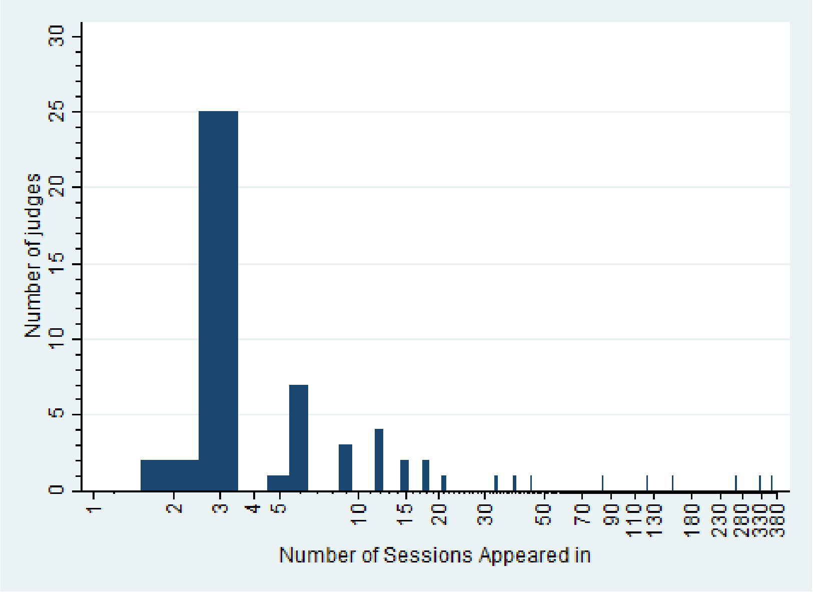
3 Preliminaries

Of the 56 judges in the show, half appeared in 5 or fewer sessions, with 9 appearing in more than 30. The minimum and maximum number of sessions appeared in by judges were respectively 2 and 360. The histogram in Figure 1 below gives the frequency distribution

[8] In the context of our empirical analysis, we thus use the term level or degree of experience as a synonym for the number of sessions a judge participated in.

of judges' participation in the show, with the number of sessions appeared in arranged in a logarithmic scale on the x-axis.

Figure 1: Frequency distribution of appearances by judges

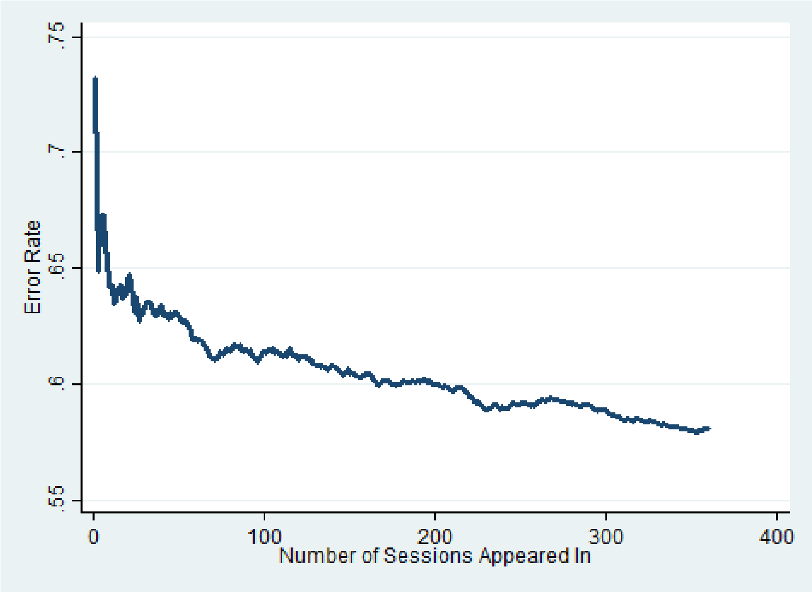


The task of any judge in any session of any episode was to determine the identity of the central character from among the group of challengers. Decisions were thus binary, either correct or erroneous. To initiate the study of trend and compare average performance with the benchmark set by random decision-making, let J be the set of judges and T_i the total number of sessions in which judge $i \in J$ appeared. If t be number of appearances (sessions), let $L(t)$ be the set of judges who appeared in no more than t sessions, and $M(t)$ be the set of judges who appeared in more than t sessions ($L(t) = \{i \in J | T_i \leq t\}$, $M(t) = \{i \in J | T_i > t\}$, for $t \geq 1$).

Next suppose that the numbers of decisions and erroneous decisions taken by judge i in his or her first t appearances are respectively $n_i(t)$ and $q_i(t)$. Let $n(t)$ and $q(t)$ be respectively the total number of decisions and total number of erroneous decisions made, across all judges, after the first t appearances. We can define $n(t) = \sum\limits_{i \in L(t)} n_i(T_i) + \sum\limits_{i \in M(t)} n_i(t)$ and $q(t) = \sum\limits_{i \in L(t)} q_i(T_i) + \sum\limits_{i \in M(t)} q_i(t)$. Then the mean error rate up to and including the t^{th} appearance is $r(t) = q(t)/n(t)$. And an individual judge's error rate over the same period is $r_i(t) = q_i(t)/n_i(t)$.

Figure 2 plots $(t, r(t))$ for all values of $t = 1, \ldots, 360$. A trend for error rate to decline with experience is apparent. We investigate this possibility formally in Section 4.

Figure 2: Mean error rate



To check whether decision quality exceeded the random benchmark, note that random decisions would generate an error rate of 0.667. With one observation per judge, his or her error rate at the end of the participation horizon $(r_i(T_i))$, the average error rate (proportion of incorrect decisions) in the sample is 0.58. It differs from the benchmark rate at the 99.9% level by a single-sample Snedecor-Cochran test (two-sided p-value = 0.0007), but only at the 90% level by a single-sample t-test (two-sided p-value = 0.0703).

4 Analysis

Figure 2 indicates a negative relationship between experience and error. An immediate explanation is learning, by which we mean any time-varying component of performance: more experience allows better understanding and recognition of habitual cues and their links to deceptive behavior, and hence produces superior judgment. As detailed earlier, other explanations are possible. It could be that later episodes, or later sessions within episodes, present easier challenger groups. Later episodes could also comprise of more able judges. The latter could arise because selection of judges into episodes is determined by performance in earlier episodes, causing judges of higher intrinsic ability to be selected more often, leading to positive correlation between ability and experience.

We now analyze if learning emerges as an explanation after either econometrically controlling for bias in the selection of judges into episodes (Section 4.1), or eliminating its possibility (Section 4.2).[9] The first approach also controls for effects arising from the chronological arrangement of challenger groups, while the second does so partially.

4.1 Learning in the full sample

We use a pooled probit regression approach to investigate if error decreased with experience. Our data are longitudinal and consitute an unbalanced panel. The model comprises two equations, one governing a judge's performance in any appearance and the other determining selection of a judge in any episode. We model performance of any judge j in any session σ as a function of the number of appearances of j in the season up to σ and other characteristics of j, characteristics of σ and the episode containing σ, and a performance characteristic of the peer judges who were present with j in σ. Selection of any judge j in any episode e is modeled as a function of j's experience and performance in the season

[9] Bias in the selection of judges into episodes can occur in two ways: producers may have incentives to re-invite certain kinds of judges, and potential judges may not randomly choose to appear. We do not have data to distinguish between such producer selection and self selection.

up to e, other characteristics of j, characteristics of e, and a performance characteristic of the peer judges who were present in e. The performance equation is as follows.

$$y^*_{i,t} = \alpha_1 + \beta_1 t + \mathbf{z}_i \boldsymbol{\gamma}_1 + \mathbf{p}_{\sigma_{i,t}} \boldsymbol{\pi}_1 + \mathbf{w}_{e_{i,t}} \boldsymbol{\delta}_1 + \theta_1 q^{e_{i,t}}_{\sigma_{i,t}} + \mathbf{e}\boldsymbol{\psi}_1 + \boldsymbol{\sigma}\boldsymbol{\phi}_1 + u^1_{i,t};\;\; y_{i,t} = 1[y^*_{i,t} > 0] \quad (1)$$

The binary dependent variable $y_{i,t}$ is the outcome for judge i in his or her t^{th} appearance, coded as 0 if the judge was not deceived, and 1 if the judge was deceived.[10] t tracks the number of appearances.

For the rest of the terms in (1), $\mathbf{z}_i$ contains time-invariant features of i, specifically the average performance of i in his or her first episode,[11] i's age on the day of entering the game show,[12] the episode number in which i appeared for the first time,[13] and i's gender. Next, letting $\sigma_{i,t}$ denote the session corresponding to i's t^{th} appearance, $\mathbf{p}_{\sigma_{i,t}}$ comprises observable characteristics of $\sigma_{i,t}$, specifically the total number of challengers, gender composition of the challenger group, and whether there is a recusal (by some judge other than i) or not. Additionally, letting $e_{i,t}$ denote the episode containing $\sigma_{i,t}$, $\mathbf{w}_{e_{i,t}}$ includes observable features specific to $e_{i,t}$: the identity of the host (standard or otherwise) and the number of sessions in the episode. Further, $q^{e_{i,t}}_{\sigma_{i,t}}$ incorporates a characteristic of the peer judges who participate in $e_{i,t}$ along with judge i. Prima facie, the decision problem for any judge in any session is individual. However, each judge was present with other judges in any episode, sometimes meeting another judge in multiple episodes. Hence a concern for relative performance could have been present. Such an incentive could alternatively have been driven by perceptions of fans' preferences over judge performance. To capture such

[10] If i recused for the t^{th} appearance, then $y_{i,t}$ is a missing value.

[11] This variable is measured by the average error rate of the judge in his or her first episode and captures unobserved performance related time-invariant individual characteristics such as intrinsic ability.

[12] Information on age was collected from publicly available sources (e.g., wikipedia). This is the only variable in our analysis for which information was obtained from outside the events of the game show as telecast.

[13] This variable is a control for potential order effects.

latent performance incentives, we use the session-level time-varying average cumulative performance of the peer group for that episode up to but not including $\sigma_{i,t}$. Finally, we use dummies to capture effects of unobserved characteristics of episodes and sessions on judge performance. The concern is that intrinsic detectability of challenger groups differs across episodes or sessions within an episode. $\boldsymbol{\sigma} = (\sigma_2, \sigma_3)$ and $\mathbf{e} = (e_2, e_3, \dots e_{145})$ represent these session and episode dummies respectively. A judge's decision is observed only when the judge is selected in an episode. We now present the selection equation.

$$s^*_{i,e} = \alpha_2 + \beta_2 x_{i,e} + \mu_2 h_{i,e} + \mathbf{z}_i \boldsymbol{\gamma}_2 + \mathbf{w}_e \boldsymbol{\delta}_2 + \eta_2 q_{i,e} + \mathbf{e}\boldsymbol{\psi}_2 + u^2_{i,e}; \quad s_{i,e} = 1[s^*_{i,e} > 0] \tag{2}$$

The binary dependent variable $s_{i,e}$ assumes value 1 (0) if judge i is selected (not selected) in episode e. $x_{i,e}$ denotes the number of days that have elapsed from the day of the episode in which judge i first appeared on the show until the end of episode $e-1$ and is a measure of experience. $h_{i,e}$ is a measure of i's performance at the beginning of e, and is a ratio which has as its denominator the number of appearances of i till the end of $e-1$. The numerator is a score, where 1 is given for every correct decision, -1 for every incorrect decision, and 0 for every recusal of i till the end of $e-1$. $x_{i,e}$ and $h_{i,e}$ serve as the exclusion restrictions for the selection equation.

For the rest of the variables, $\mathbf{z}_i$ and $\mathbf{w}_e$ refer respectively to time-invariant judge-specific regressors and episode-specific regressors, and contain respectively the same regressors as for equation (1). $q_{i,e}$ is a performance measure of the peer judges present in episode e, and is the cumulative average performance of these judges up to the end of episode $e-1$. We include this variable to account for any group aspect of the process driving selection of judges into episodes. Lastly, $\mathbf{e}$ contains the episode dummies, as defined for (1).

The estimation sample comprises two or three observations per judge for every episode in which the judge participated (recall that the first six episodes of the show contain two sessions each while the rest consists of three sessions). It includes one observation per judge for every episode in which a judge did not participate. This process leads to 9239 observations after recusals are accounted for (see III in Section 2). The construction of the data requires the imputation of values of variables for judges corresponding to episodes in which they did not appear. The choices faced and the rules used for construction are detailed in Appendix B.

We assume that $u^1_{i,t}, u^2_{i,t} | t, \mathbf{z}_i, \mathbf{p}_{\sigma_{i,t}}, \mathbf{w}_{e_{i,t}}, q^{e_{i,t}}_{\sigma_{i,t}}, x_{i,e}, h_{i,e}, q_{i,e}, \mathbf{e}, \boldsymbol{\sigma} \sim N(0, 0, 1, 1, \rho)$, where $\rho = corr(u^1, u^2)$. Also, conditional errors are allowed to be arbitrarily serially correlated.[14] Further, our model controls for unobserved heterogeneity stemming from differences in intrinsic ability across judges (see fn. 11), and from session and episode level factors. It is therefore unlikely that $u^1_{i,t}$ or $u^2_{i,t}$ contains unobserved individual-specific outcome or selection related factors leading to endogeneity.

We estimate this partial likelihood model under joint normality of (u^1, u^2). Given the assumptions of the model, partial MLE is both consistent and asymptotically efficient.[15] Cluster-adjusted standard errors are used for inference, with clustering by judge.[16] The output is presented in Table 1. Names of independent variables appearing in all tables of estimates in the paper are collected together in Appendix C, where they are described and defined.

The estimates for equation (1) show that lower error in the first appearance episode lowers overall error. Hence intrinsic ability affects performance. Further, the correlation coefficient (ρ) between equations (1) and (2) is significant at the 90% level, weakly suggesting that the equations are not independent, implying prior performance possibly influences the selection of judges into episodes.[17] A key finding is that session dummies

[14] Conditional error terms in all probit models in the paper, described in equations (1) through (3), are assumed to be standard normal and allowed to be arbitrarily serially correlated.

[15] See e.g., Wooldridge (2010), Semykina and Wooldridge (2018) for details.

[16] All regression models reported in the paper implement cluster correction, with clustering at the level of the judge.

[17] However, the product moment correlation coefficient between *Performance in first episode* and the number of episodes appeared in for judges was 0.0326 and it was insignificant (p-value = 0.8114), sug-

Table 1: Experience and learning: joint pooled probit estimates

	Performance equation (1)		
	Dependent variable: judge decision (0 if correct)		
Focal regressor (t)	Appearance number	-0.00249***	(0.00067)
Time-invariant judge-specific characteristics ($\mathbf{z}_i$)	Performance in first episode	0.48737**	(0.16098)
	Episode in which first appeared	-0.00198	(0.00200)
	Age in days on first appearance	0.00004**	(0.00001)
	Gender (1 if male)	-0.19964**	(0.07193)
Session level characteristics ($\mathbf{p}_{\sigma_{i,t}}$)	Number of female challengers	0.06833**	(0.02527)
	Total number of challengers	0.23496***	(0.05638)
	Recusal (1 if recusal)	-0.41231	(0.38369)
Episode level characteristics ($\mathbf{w}_{e_{i,t}}$)	Host (1 if standard)	0.25269	(0.49477)
	Number of sessions in episode	-0.24353	(0.57323)
Peer effect ($q^{e_{i,t}}_{\sigma_{i,t}}$)	Average peer cumulative performance	-0.01240	(0.69188)
Session dummies ($\boldsymbol{\sigma}$)	Session 2	-0.04068	(0.10770)
	Session 3	-0.07095	(0.05223)
Episode dummies ($\boldsymbol{e}$)	Yes		
	Selection equation (2)		
	Dependent variable: judge selection (1 if selected)		
Exclusion restrictions	Experience till previous episode ($x_{i,e}$)	0.00392***	(0.00033)
	Self cumulative peformance ($h_{i,e}$)	-1.28870	(0.76851)
Time-invariant judge-specific characteristics ($\mathbf{z}_i$)	Performance in first episode	0.51720	(0.41517)
	Episode in which first appeared	-0.00411	(0.00317)
	Age in days on first appearance	0.00004	(0.00004)
	Gender (1 if male)	-0.33711	(0.34307)
Episode level characteristics ($\mathbf{w}_{e_{i,t}}$)	Host (1 if standard)	-5.89751***	(1.63467)
	Number of sessions in episode	-0.56489	(0.34895)
Peer effect ($q_{i,e}$)	Lagged average peer cumulative performance	9.24015***	(2.76535)
Episode dummies ($\boldsymbol{e}$)	Yes		
Observations	9239		
Correlation coefficient (ρ)	-0.10854		
p-value	0.06025		
Elasticity	-0.19379		

Cluster adjusted standard errors in parentheses. ** $p < 0.01$, *** $p < 0.001$.

For definitions of independent variables, see Appendix C.

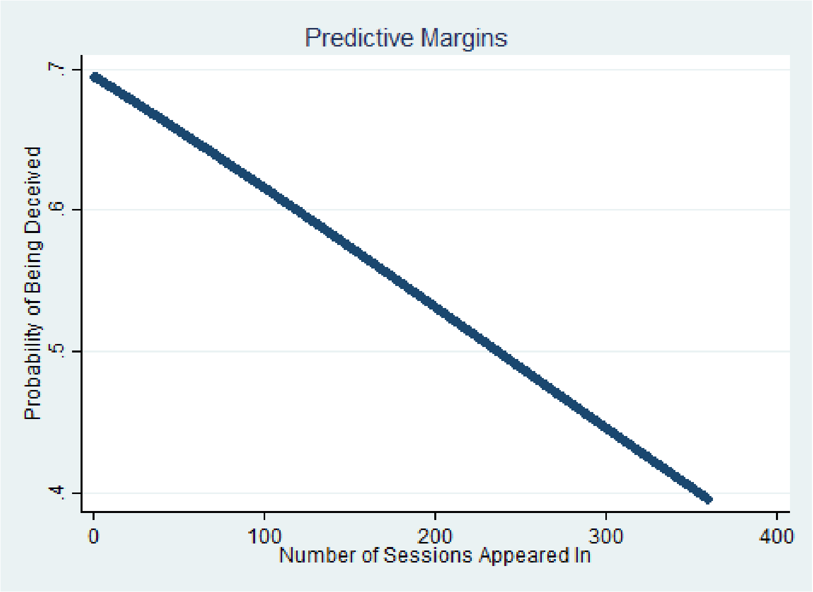
are insignificant, implying that unobserved characteristics of sessions, including challenger groups, have no impact on performance. Hence we find no evidence that later sessions systematically involved easier challenger groups. Additionally, we find that episode dummies, which we do not report for brevity, are usually insignificant for (1), with only 29 achieving significance. Negatively significant dummies, indicative of the presence of easier challenger groups, occur mostly in the first half of the season. Hence we find no evidence that later episodes systematically involved easier challenger groups. The central finding for the performance equation in Table 1 is that *Appearance number* is negative and significant. A conjoint finding for the selection equation is that neither prior performance nor intrinsic ability influence selection. These findings hence lead us to conclude that experience positively influences performance in our data, after controlling for effects arising from chronological arrangements of groups of challengers and judges. The force is not negligible, as the elasticity calculation shows that a doubling of the level of experience causes the probability of error to decline by more than 19%.[18] Figure 3 shows the corresponding relationship between experience and the predicted probability of error.

The estimates for equation (1) produce additional findings. First, there appears to be an effect of age on performance, as *Age in days on first appearance* is significant, implying higher error may be produced by higher age ceteris paribus. Second, gender appears to affect outcome. Specifically, *Gender* is significant, suggesting that male judges had lower error in detection, as is *Number of Female Challengers*, suggesting that female challengers were more successful at deception. And third, couples appear more successful at deception than individuals, as *Total number of challengers* is significant (see II in Section 2). However, couples appeared as challengers in only 4 out of 429 sessions, leading

gesting that selection bias is not a strong force in our sample.

[18] Elasticity is defined as $\frac{dy}{dx}.\frac{x}{y}$, where x and y are respectively the independent and dependent variables, and is the point elasticity averaged over the sample of the x-variable.

Figure 3: Experience and the estimated probability of error



to limited variation in the values of the variable.[19]

As a robustness check, we estimated equations (1) and (2) using the two-step procedure (Heckman 1979; Semykina and Wooldridge 2018), whereby equation (2) is estimated first, and an augmented version of equation (1) is estimated subsequently. The augmentation is due to the presence of the inverse Mills ratio (IMR) generated from the first step as an additional explanatory variable. We describe these results, but do not report them for brevity. These estimates continued to provide null evidence in favor of later episodes or later sessions within episodes involving easier challenger groups. With analytical standard errors, the focal regressor continued to be strongly significant, reinforcing the finding of learning, but the IMR was significant as well, suggesting performance trend may be additionally affected by selection bias. There was no evidence of such selection influences with bootstrapped standard errors (p-value of IMR = 0.195). However, the focal regressor

[19] By contrast, there was substantial variation in values of the variables capturing gender. Half the judges in any session were male, and about 60% and 36% of challenger groups (out of the 425 where challengers were individuals) were all male and all female respectively.

was only weakly significant (p-value = 0.062) in this case.

To summarize, first, we find no evidence to suggest that later episodes or later sessions within episodes involved easier challenger groups. In particular, session dummies are always insignificant. Second, there was mixed evidence with respect to the presence of selection bias which affected performance trend. The main joint estimation procedure provided weak support for the possibility, while the two-step procedure with analytical standard errors provided strong support. Other tests provided no support. Finally, the main estimates strongly showed the presence of learning. Evidence in favor of learning emerging from the two-step procedure was strong with analytical standard errors but weak with bootstrapped standard errors.

The absence of unanimity with respect to strong significance of the focal regressor across the procedures prompts us to search for further evidence in favor of learning. The approach in the current section econometrically controlled for bias in the selection of judges into episodes, imposing a specific model of selection. We hence explore an alternative in Section 4.2 which obviates the need to account for such selection bias.

Before ending, we note that the analysis in Section 4 proceeds by pooling observations across judges. We also investigated learning at the level of individual judges, by estimating a version of equation (1) reduced by dropping time-invariant judge-specific variables ($\mathbf{z}_i$) for each judge. Significant learning did not emerge for any judge. We thus do not find evidence for learning at the level of individual judges in our sample.

4.2 Intra-episode learning

We recall that episodes in the show contained multiple sessions and that judges were fixed for any given episode. Assuming no interaction between time-invariant and time-dependent components of performance, this implies that any change in performance over the course of an episode cannot be attributed to bias in the selection of judges. Hence, an analysis restricted to intra-episode data can proceed without needing to account for selection of judges. However, the variable representing experience in such an analysis would be collinear with the session dummies, implying the absence of control for unobserved session level characteristics, particularly how easy or difficult a challenger group is. We appeal here to the finding from Section 4.1 indicating there was no tendency for unobserved session level characteristics to affect performance. This finding provides the foundation for the analysis of this section.

The simplest way forward is to focus on the first appearance episode for each judge, where performance-dependent selection is by definition absent. In Section 4.2.1, we analyze intra-episode learning, focussing on data from the first appearance episode for each judge. In Section 4.2.2, we extend our approach to account for data from all episodes.

4.2.1 Learning over the episode of first appearance

In this section, we analyze intra-episode learning taking data only from the first appearance episode for each judge. First appearance episodes of 7 judges had only 2 sessions, with 3 sessions for the remaining 49 judges, leading to a maximum of 161 observations. One judge recused once in the first episode of appearance, leaving 160 observations. The model of performance is given in (3) below.

$$y^*_{i,t} = \alpha_3 + \beta_3 t + \mathbf{z}_i \boldsymbol{\gamma}_3 + \mathbf{p}_{i,t} \boldsymbol{\pi}_3 + \mathbf{w}_i \boldsymbol{\delta}_3 + \theta_3 q_{i,t} + u^3_{i,t}; \;\; y_{i,t} = 1[y^*_{i,t} > 0] \tag{3}$$

$y_{i,t}$ is the outcome for judge i in his or her t^{th} session, as in (1) (as before, coded as 0 if the judge was not deceived, and 1 if the judge was deceived). The main regressor, t, is as defined for (1). For judges whose first episodes of appearance had 3 sessions, t can take values 1, 2 or 3. For the other judges, t can take values 1 or 2.[20]

For the rest of the terms in (3), $\mathbf{z}_i$ is as defined in (1), except that average performance of i in his or her first episode is dropped, as this variable is now perfectly correlated with

[20] We also estimated the model assuming t can take values 1 or 3 for judges whose first episode of appearance had 2 sessions. No qualitative differences in results were detected, and we do not report these estimates for brevity.

the dependent variable. Next, $\mathbf{p}_{i,t}$ is identical to $\mathbf{p}_{\sigma_{i,t}}$, $\mathbf{w}_i$ is identical to $\mathbf{w}_{e_{i,t}}$, and $q_{i,t}$ is identical to $q^{e_{i,t}}_{\sigma_{i,t}}$, as these are defined in (1), with the restriction that reference is only to the first appearance episode for any i.

Note that equation (3) does not contain episode dummies. However, control is instituted for unobserved episode level characteristics through the variable *Episode in which first appeared* (an element of $\mathbf{z}_i$), as appearing in Table 1. This variable controlled for order effects in equations (1) and (2), an issue of no relevance here as attention is restricted to a single episode, rendering it a pure episode index variable able to capture the effect of unobserved characteristics. We also estimated a version of equation (3) which replaces this episode index with the set of relevant episode dummies. Pooled probit estimation results with the index variable are in the first column of Table 2 while those with the dummy variables are in the second column.

Both columns display a negative sign for the focal regressor. The coefficient is significant in the case of the estimation with the index variable. It is insignificant at the 95% level when dummy variables are used instead. However, significance in this case is achieved at the 90% level (p-value = 0.074), providing weak evidence of learning. This fall in the significance level could be due to the higher standard errors generated through loss in the degrees of freedom. Figure 4 shows the relationship between the level of experience and the predicted probability of error, corresponding to the first column of Table 2.[21]

A problem with the above approach is that omitting the control for unobserved judge characteristics (average performance in the first appearance episode) may cause unobserved individual heterogeneity to be correlated with observed individual-specific regressors such as the episode number of first appearance, gender etc. An alternative would be to estimate (3) using correlated random effects (Chamberlain 1980), which incorporates judge level averages of all time-varying regressors as additional controls (Mundlak 1978) to proxy individual-level unobserved heterogeneity. Results are presented in the third

[21] Table 2 reports a semi elasticity, defined as $\frac{dy}{dx}.\frac{1}{y}$, where x and y are respectively the independent and dependent variables, as the independent variable *Session* was treated as discrete in our regression.

Table 2: Intra-episode learning: first episode of appearance

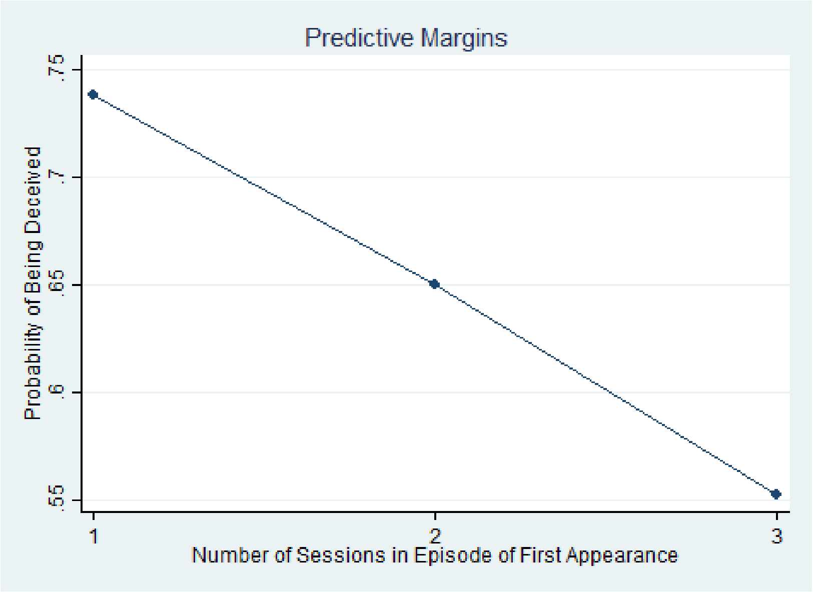
		Dependent variable: judge decision (0 if correct)		
		pooled probit (episode index)	pooled probit (episode dummies)	correlated random effects
Focal	Session	-0.26611*	-0.35760	-0.25996*
regressor (t)		(0.12514)	(0.20001)	(0.12495)
Time	Episode in which	0.00113	-	0.00265
invariant	first appeared	(0.00342)	(-)	(0.00364)
judge	Age in days on	0.00001	-0.00040***	0.00002
specific	first appearance	(0.00004)	(0.00010)	(0.00004)
characteristics	Gender	-0.41848	-0.76087	-0.47715
$(\mathbf{z}_i)$	(1 if male)	(0.25844)	(0.51680)	(0.27333)
Session level	Number of	-0.04232	-0.09556	-0.06342
	female challengers	(0.08079)	(0.13498)	(0.08381)
characteristics	Total number	1.00399***	1.26159***	1.07998***
	of challengers	(0.13242)	(0.19282)	(0.25059)
$(\mathbf{p}_{\sigma_{i,t}})$	Recusal	-1.17168	-33.59296	-0.84677
	(1 if recusal)	(0.71291)	(.)	(0.44135)
Episode	Host	0.49018	-38.67860***	-0.11284
level	(1 if standard)	(0.39805)	(0.87498)	(0.43307)
characteristics	Number of sessions	0.48347	0.81918	0.07839
$(\mathbf{w}_{e_{i,t}})$	in episode	(0.28279)	(1.03288)	(0.74570)
Peer effect	Average peer	-0.91036	-3.72099	3.80848
$(q_{i,t})$	cum. perf.	(1.73334)	(4.33897)	(4.26399)
Judge averages		-	-	Yes
Episode dummies		-	Yes	-
Observations†		156	156	156
(Semi) Elasticity		-0.15356	0	-0.15017

†Values for the peer effect variable are not defined for the first session of the first episode of the show. We hence lose 4 observations from 160, leaving us with a total of 156.

Cluster adjusted standard errors in parentheses. * $p < 0.05$, *** $p < 0.001$.

For definitions of independent variables, see Appendix C.

Figure 4: Experience and error probabilities: first episodes of appearance



column of Table 2 and provide support for learning. We conclude that there is sufficient evidence to suggest the presence of learning during the first episode of appearance in our sample, indicating learning by novices can take place under short horizons.

4.2.2 Average intra-episode learning

In this section, we extend the approach of Section 4.2.1 to account for all episodes, while continuing to ignore information arising from inter-episode variation in performance. The question we pursue is whether there is any evidence for performance improvement over the course of an average episode, considering all appearance episodes together.

For this purpose, we represent each judge by the averages of her performances for the first, second and perhaps third sessions, across all episodes of appearance. For any episode j with three sessions, let j_1, j_2, and j_3 be respectively its first, second, and third sessions and let j_1 and j_2 be its first and second sessions if it has only two. Consider a judge i who appears in $T_i \geq 1$ episodes. Let the first $T_i^2 \geq 0$ episodes contain only two sessions, with the remaining $T_i^3 \geq 0, T_i^2 + T_i^3 = T_i$, containing three sessions.[22] Denote performance by y (0 if not deceived, 1 otherwise), with the observation missing in the event of recusal. Suppose $n_{i,k}$ is the number of decisions taken by i over all k^{th} sessions in the episodes in which he or she has appeared, and y_{i,j_k} is i's performance in the k^{th} session of the j^{th} episode of appearance. Then, for k=1,2, i's average performance for the k^{th} session over all appearance episodes is $y'_{i,k} = (1/n_{i,k}) \sum_{j=1}^{T_i} y_{i,j_k}$.

If $T_i^3 = 0$, i is represented by two observations, $y'_{i,1}$ and $y'_{i,2}$. If $T_i^3 > 0$, i's average performance for the 3rd session over all appearance episodes is $y'_{i,3} = (1/n_{i,3}) \sum_{j=T_i^2+1}^{T_i} y_{i,j_3}$, where $n_{i,3}$ and y_{i,j_3} respectively follow from $n_{i,k}$ and y_{i,j_k} above, for $k = 3$. i is then represented by three observations, $y'_{i,1}$, $y'_{i,2}$ and $y'_{i,3}$.

The values of $y'_{i,1}$, $y'_{i,2}$ and $y'_{i,3}$ are between 0 and 1, end-points included, for all judges, whenever defined.[23] 3 judges appeared only in episodes with exactly two sessions each, all other judges appearing in at least one episode with exactly three sessions. We hence have 165 observations. The following model is used for performance:

$$y'_{i,t} = \alpha_4 + \beta_4 t + \mathbf{z}_i \boldsymbol{\gamma}_4 + u_{i,t}^4 \tag{4}$$

The model is estimated using pooled OLS. t can take values 1 or 2 in (4) if i participated only in episodes with exactly two sessions, and values 1, 2 or 3 otherwise.[24] $\mathbf{z}_i$ is as defined for (1). We only use time-invariant individual-specific regressors as independent variables. We do not utilize session or episode level variables as controls as these would then have to be used in averaged form, which does not seem to yield decision-relevant statistics.

Performance improves over the course of an average episode, as constructed above, if β_4 estimated from this model is negative. Results are given in Table 3. Directionally, we find improvement in performance. Significance is not achieved at the 95% level. How-

[22] This partition follows from only the early episodes having exactly two sessions (see Section 2).

[23] $y'_{i,1}$ and $y'_{i,2}$ were defined for all judges. $y'_{i,3}$ was defined for all judges who participated in at least one episode with three sessions.

[24] Alternative estimates were generated using the procedure described in fn. 20. No qualitative difference in results were detected, and we do not report those estimates for brevity.

ever, the coefficient is marginally insignificant at that level (p-value = 0.051), suggesting weak evidence of performance improvement and hence intra-episode learning. The only other variable having explanatory power is *Performance in First Episode*, suggesting as in Section 4.1 that intrinsic ability affects performance.

Table 3: Intra-episode learning: all episodes

	Dependent variable: average judge decision	
Focal	Session	-0.06676
regressor (t)		(0.03417)
Time-invariant	Performance in	0.67211***
	first episode	(0.07624)
judge-specific	Episode in which	-0.00012
	first appeared	(0.00045)
characteristics	Age in days on	0.00001
	first appearance	(0.00001)
($\mathbf{z}_i$)	Gender	-0.02073
	(1 if male)	(0.04144)
Observations	165	

Cluster adjusted standard errors in parentheses. *** $p < 0.001$.

For definitions of independent variables, see Appendix C.

5 Conclusions

The need to identify deceptive or opportunistic individuals may arise in many economic situations. Existing evidence is somewhat bleak and suggests that humans are not necessarily highly skilled at deception detection. The literature has therefore searched for valid indicators of deception and factors which may induce improved detection.

It has been conjectured that experience may be a factor favoring improved detection. However, prior research has not found strong support for the hypothesis that individuals with greater exposure to situations with potential deception can learn from such experiences and augment their ability to detect it. We pursue this hypothesis by analyzing data constructed from events in a TV game show. Our evidence suggests the presence of learning, with performance improving through experience, perhaps because judges have more personal data on decision-making in analogous situations.

Individuals with no special training can thus learn to better filter misinformation through experience. Policy attempting to control misinformation should hence encourage learning and account for different levels of experience of potential receivers.

A question that remains unanswered in our investigation is what it is that is learnt. Prior research has pointed to various cues that may indicate deception. Our approach is aggregative and does not isolate particular cues that individuals may be learning through experience. This limitation follows to an extent from the show not using special equipment that could have tracked cues emerging from the body or voice tones. Semantic or syntactic analysis of oral communication, with a view to understanding if verbal or linguistic cues were learnt, is however feasible. We plan to pursue this in future research.

We conduct our examination in the domain of face-to-face oral communication. Deception is possible in other domains as well, such as ones with written communication. It would be of interest to know if experience is valuable in such domains. Also, our sample of judges is restricted, mostly being from the world of entertainment. We leave the question of whether learning is feasible in more general samples for future research. Further, deception is always present in our environment, whereas in reality the detection of deception may have to account for the possibility of its absence. Our results cannot comment on the interesting question of whether experience is useful in such settings. Additionally, each judge, while an independent decision-maker, was always in the company of other judges in the show. Whether learning is possible with isolated decision-makers may be worth pursuing, particularly in light of recent findings regarding deception detection with group decision-making (Klein and Epley 2015).

References

[1] Bak-Coleman, J., Alfano, M., Barfuss, W., Bergstrom, C., Centeno, M., Couzin, I., Donges, J., Galesic, M., Gersick, A., Jacquet, J., Kao, A., Moran, R., Romanczuk, P., Rubenstein, D., Tombak, K., Van Bavel, J. and Weber, E. (2021): "Stewardship of global collective behavior," *Proceedings of the National Academy of Sciences of the U.S.A.*, 118, e2025764118.

[2] Becker, G. (1975): *Human Capital.* National Bureau of Economic Research, New York.

[3] Belot, M., Bhaskar, V. and van de Ven, J. (2012): "Can observers predict trustworthiness?" *Review of Economics and Statistics*, 94, 246-59.

[4] Belot, M. and van de Ven, J. (2017): "How private is private information? The ability to spot deception in an eeconomic game," *Experimental Economics*, 20, 19-43.

[5] Brosig, J. (2002): "Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game," *Journal of Economic Behavior and Organization*, 47, 275-90.

[6] Chamberlain, G. (1980): "Analysis with qualitative data," *Review of Economic Studies*, 47, 225-38.

[7] Crawford, V., and Sobel, J. (1982): "Strategic information transmission," *Econometrica*, 50, 1431-51.

[8] Darai, D., and Grätz, S. (2010): "Determinants of successful cooperation in a face-to-face social dilemma," University of Zurich WP # 1006.

[9] DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K. and Cooper, H. (2003): "Cues to deception," *Psychological Bulletin*, 129, 74-118.

[10] DePaulo, B. and Pfeifer, R. (1986): "On-the-job experience and skill at detecting deception," *Journal of Applied Social Psychology*, 16, 249-67.

[11] Ekman, P. and O'Sullivan, M. (1991): "Who can catch a liar?" *American Psychologist*, 46, 913-20.

[12] Elaad, E. (2003): "Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to tell lies," *Applied Cognitive Psychology*, 17, 349-63.

[13] Erev, I., and Haruvy, E. (2016): "Learning and the economics of small decisions," in Kagel, J. and Roth, A. eds. *The Handbook of Experimental Economics, vol. 2.* Princeton Univerity Press, Princeton, New Jersey.

[14] Fujiwara, T., Müller, K. and Schwartz, C. (2021): "The effect of social media on elections: Evidence from the United States," NBER WP 28849.

[15] Gertner, R. (1993): "Game shows and economic behavior: Risk-taking on *Card Sharks*," *Quarterly Journal of Economics*, 108, 507-21.

[16] George, J., Marett, K., Burgoon, J., Crews, J., Cao, J., Lin, M. and Biros, D. (2004): "Training to detect deception: An experimental investigation," *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, v. 1 (doi:10.1109/HICSS.2004.1265082).

[17] Gokhmann, S., Hancock, J., Prabhu, P., Ott, M. and Cardie, C. (2012): "In search of a gold standard in studies of deception," in Fitzpatrick, E., Bachenko, J. and Fornaciari, T. eds. *Proceedings of the EACL Workshop on Computational Approaches to Deception Detection.*

[18] Hartwig, M. and Bond, C. (2011): "Why do lie-catchers fail? A lens model meta-analysis of human lie judgements," *Psychological Bulletin*, 137, 643-59.

[19] Hauch, V., Sporer, S., Michael S. and Meissner, C. (2016): “Does training improve the detection of deception? A meta-analysis,” *Communication Research*, 43, 283-343.

[20] Heckman, J. (1979): “Sample selection bias as a specification of error,” *Econometrica*, 47, 353-61.

[21] Institutional Investor (2019): “Fake news creates real losses,” (`https://www.institutionalinvestor.com/article/b1j2ttw22xf7n6/Fake-News-Creates-Real-Losses`).

[22] Kim, H. K., Ahn, J., Atkinson, L. and Kahlor, L. E. (2020): “Effects of covid-19 misinformation on information seeking, avoidance, and processing: A multicountry comparative study,” *Science Communication*, 42, 586-615.

[23] Klein, N. and Epley, N. (2015): “Group discussion improves lie detection,” *Proceedings of the National Academy of Sciences of the U.S.A.*, 112, 7460-5.

[24] Kraut, R. and Poe, D. (1980): “On the line: The deception judgements of customs inspectors and laymen,” *Journal of Personality and Social Psychology*, 36, 380-91.

[25] Maranto, C. and Rodgers, R. (1984): “Does work experience increase productivity? A test of the on-the-job training hypothesis,” *Journal of Human Resources*, 19, 341-57.

[26] Masip, J., Martinez, C., Blandón-Gitlin, I., Sánchez, N., Herrero, C. and Ibabe, I. (2018): “Learning to detect deception from evasive answers and inconsistencies across repeated interviews: A study with lay respondents and police officers,” *Frontiers in Psychology*, 8, a. 2207.

[27] Mincer, J. (1974): *Schooling, Experience and Earnings.* National Bureau of Economic Research, New York.

[28] Mundlak, Y. (1978): “On the pooling of time series and cross section data,” *Econometrica*, 46, 69-85.

[29] National Association of Colleges and Employers (2012): *Salary trends through salary survey: A historical perspective on starting salaries for new college graduates*, available at https://www.naceweb.org/job-market/compensation/salary-trends-through-salary-survey-a-historical-perspective-on-starting-salaries-for-new-college-graduates/.

[30] Newell, A., and Rosenbloom, P. (1981): “Mechanisms of skill acquisition and the law of practice,” in Anderson, J. eds. *Cognitive Skills and their Acquisition.* Erlbaum, Hillsdale, New Jersey.

[31] Ockenfels, A. and Selten, R. (2000): “An experiment on the hypothesis of involuntary truth-signalling in bargaining,” *Games and Economic Behavior*, 33, 90-116.

[32] Park, H., Levine, T., McCornack, S., Morrison, K. and Ferrara, S. (2002): “How people really detect lies,” *Communication Monographs*, 69, 144-57.

[33] Pfundmair, M., Erk, W. and Reinelt, A. (2017): “Lie to me - Oxytocin impairs lie detection between sexes,” *Psychoneuroendocrinology*, 84, 135-8.

[34] Rose, J. (2017): “Brexit, trump, and post-truth politics,” *Public Integrity*, 19, 555-8.

[35] Semykina, A. and Wooldridge, J. (2018): “Binary response panel data models with sample selection and self selection,” *Journal of Applied Econometrics*, 33, 179-97.

[36] Serra-Garcia, M. and Gneezy, U. (2021): ‘Mistakes, overconfidence, and the effect of sharing on detecting lies,” *American Economic Review*, 111, 3160-83.

[37] Taylor, M. (2001): “Self-employment and windfall gains in Britain: Evidence from panel data,” *Economica*, 68, 539-65.

[38] Turmunkh, U., van den Assem, M., and van Dolder, D. (2019): "Malleable lies: Communication and cooperation in a high stakes TV game show," *Management Science*, 65, 4795-812.

[39] Vrij, A. and Mann, S. (2001): "Who killed my relative? Police officers' ability to detect real-life high-stake lies," *Psychology, Crime and Law*, 7, 119-32.

[40] Wang, J., Spezio, M., and Camerer, C. (2010): "Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games," *American Economic Review*, 100, 984-1007.

[41] Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data.* MIT Press, Cambridge, Massachusetts.

[42] Zuckerman, M., DePaulo, B. and Rosenthal, R. (1981): "Verbal and nonverbal communication of deception," *Advances in Experimental Social Psychology*, 14, 1-59.

[43] Zuckerman, M., Koestner, R. and Alton, A. (1984): "Learning to detect deception," *Journal of Personality and Social Psychology*, 46, 519-28.

Appendix A: Details of events within a regular session

Here we briefly and chronologically describe all events within any regular session. Any session begins with the camera focusing on the three challengers. The host then asks every challenger, 'What is your name, please?', following which each replies, 'My name is [central character's name].' The host then reads aloud a signed affidavit about the central character, as the judges read their copies of the same affidavit. The judges are then allowed to ask questions to the challengers one at a time, summoning them by their numbers (numbers 1, 2, or 3). Every judge has an allotted time, and the host rings a bell once this time is up. Upon completion of the question-answer rounds, the host asks the judges to determine which challenger they think is the central character. No communication is allowed between any party at this stage. Consultation between parties (except possibly with the host) is not allowed at any stage. Each judge has to write the respective challenger number on a card and reveal it to the host. Once all judges have written on their cards, the results are revealed publicly. The host then asks, ‘Will the real [central character’s name] please stand up?’, at which point the true central character reveals himself/herself. The impostors also introduce themselves with their real names. This signals the end of the session, unless the challengers are interrogated on issues discussed during the question-answer phase, which happened occasionally.

Appendix B: Dataset for systems estimation by maximum likelihood

As mentioned in Section 5.1, the estimation sample comprises two or three observations per judge for every episode in which the judge participated (recall that the first six episodes of the show contain two sessions each while the rest consists of three sessions). It includes one observation per judge for every episode in which a judge did not participate. There are 9256 observations altogether, of which 17 observations are excluded from the sample due to recusals (see III in Section 3). Data on time-invariant judge and episode-specific regressors are available for all judges and episodes. Data on session-specific regressors (including the focal regressor of the performance equation) are available for all sessions of an episode in which a judge had participated.

Equation (1) contains a peer-specific regressor. Let the session in which judge i makes his/her t^{th} appearance be $\sigma_{i,t}$. Let the episode that contains this session be $e_{i,t}$. The peer group’s average cumulative performance in $\sigma_{i,t}$ (in the performance equation) measures the ratio of the total number of incorrect decisions of i’s peers to the total number of appearances made by i’s peers till the end of session $\sigma_{i,t} - 1$ (which may not necessarily belong to episode $e_{i,t}$).

In constructing the dataset, missing values need to be treated specially. To incorporate one observation per judge for every episode in which a judge did not participate, we generate missing values for some regressors in those episodes as described below. This requires imputation of values of these regressors for those episodes. This is a non-trivial issue for the following regressors and requires making some assumptions.

The regressors exclusive to the selection equation, viz., $h_{i,e}$ and $x_{i,e}$ are defined at the episode level. First, consider the cumulative performance of judge i at the beginning of episode e, viz. $h_{i,e}$. Suppose judge i makes his/her first appearance in the e_0^{th} episode. We replace the missing values for the judge's cumulative performance in all episodes up to $e_0 - 1$ with 0. Note, however, that given our scoring rule, a zero cumulative performance could also result if the judge had made an equal number of correct and incorrect decisions till episode $e_0 - 1$. The underlying assumption is that a zero value of cumulative performance does not provide any information about the judge's track record or help predict her future performance. We test if results are robust to this convention by using various alternatives to 0 for these missing values, including -0.34, which corresponds to the random choice benchmark. Qualitative results mostly remained unchanged across these rules. In particular, *Appearance Number* remained negative and significant in the performance equation, suggesting, as before, that experience enables learning and thereby yields performance improvement.

Second, the number of days that have elapsed since the day a judge had appeared in the show for the first time until the end of episode $e-1$, viz. $x_{i,e}$, has missing values for a judge until a judge's second episode of appearance. We replace these missing values with 0 and assign the value 1 to $x_{i,e}$ in a judge's second episode of appearance.

For $q_{i,e}$, note that existence of peers in an episode is contingent on selection in that episode. We emulate peers for non-participants in episode e by calculating the average cumulative performance, computed up to the end of episode $e-1$, of judges who actually participated in episode e. The idea is that if i had been chosen in that episode, then i would have randomly replaced one of the four judges who actually participated in it.

Appendix C: List of regressors

Here we list all independent variables found in the tables of estimates reported in the paper, across all such tables. The list is presented in alphabetical order.

- *Age in Days on First Appearance*: The age of a judge, measured in days, on the day of the episode in which the judge participated for the first time.

  Dates of birth of judges were collected from public sources such as wikipedia.

- *Appearance Number*: The total number of sessions a judge has participated in, including current session.

  This variable is denoted by t in (1), (3) and (4).

- *Average Peer Cumulative Performance*: Suppose judge i has 3 peers in episode e, P_1, P_2 and P_3. Let e contain some session σ. Let $\tilde{A}_{k,\sigma}$ and $\tilde{M}_{k,\sigma}$ be respectively the number of appearances and number of incorrect decisions of P_k $(k = 1, 2, 3)$ until and including session $\sigma - 1$. Then the value of the variable for session σ is defined as $\sum_{k=1}^{3} \tilde{M}_{k,\sigma} / \sum_{k=1}^{3} \tilde{A}_{k,\sigma}$.

- *Episode in which First Appeared*: The serial number of the episode in which a judge made his or her first appearance in the show.

- *Experience till previous episode*: The number of days that have elapsed since the day the judge first appeared in the show till the end of the previous episode.

  For the definition in episodes prior to the second appearance episode for any judge, see Appendix B.

- *Gender*: Dummy variable indicating the sex of a judge (1 if male).

  Gender classification was as announced by the host.

- *Host*: Dummy variable to indicate if the host is different from the usual one (1 for standard host).

- *Lagged average peer cumulative performance*: Suppose judge i has 3 peers in episode e, P_1, P_2 and P_3. Let $\tilde{A}_{k,e}$ and $\tilde{M}_{k,e}$ be respectively the number of appearances and number of incorrect decisions of P_k ($k = 1, 2, 3$) until and including episode $e - 1$. Then the value of the variable for episode e is defined as $\sum_{k=1}^{3} \tilde{M}_{k,e} / \sum_{k=1}^{3} \tilde{A}_{k,e}$.

  For the definition in episodes in which i did not participate, see Appendix B.

- *Number of Female Challengers*: The total number of female challengers in a session.

- *Number of Sessions in Episode*: The total number of sessions in a particular episode (2, 3).

- *Performance in First Episode*: Suppose judge i took n_i decisions in his or her first episode; then performance in first episode is $(1/n_i) \sum_{\tau=1}^{n_i} y_{i,\tau}$, where $y_{i,\cdot}$ is as defined for (1).

- *Recusal*: Dummy variable indicating if there was a recusal in a session (1 if recusal).

- *Self Cumulative Performance*: Suppose judge i earns 1 point for a correct decision, -1 point for an incorrect decision, and 0 for a recusal. Then self cumulative performance of i at the beginning of episode e is the ratio of the total points earned till the end of $e - 1$ to the total number of appearances till the end of $e - 1$.

  For the definition in episodes in which i did not participate, see Appendix B.

- *Session*: The serial number of a session within the episode.

- *Total Number of Challengers*: The total number of challengers in a session.