Behavior Research Methods
https://doi.org/10.3758/s13428-020-01526-w

The Difference Spotting Task: A new nonverbal measure of cheating behavior

Jinting Liu[1,2,3] · Qiang Shen[4,5] · Jieting Zhang[1,2,3] · Urielle Beyens[1,2,3] · Wei Cai[6,7] · Jean Decety[8,9] · Hong Li[1,2,3]

Accepted: 15 December 2020

Abstract

To understand when, how, and why people cheat, the ability to detect cheating in a laboratory setting is crucial. However, commonly used paradigms are confronted with a conflict between allowing participants to believe they can cheat unnoticed and allowing experimenters to detect cheating. This project aimed to develop and establish a new nonverbal task to resolve this conflict. Study 1 and Study 2 developed a new unsolvable paradigm called the Difference Spotting Task. In Study 1, participants were incentivized to indicate whether they found any difference between a pair of pictures without being asked to point the difference(s) out, so they could overreport their performance to earn extra money. Unbeknownst to them, the pairs of pictures from half of the items were identical so that the task could not be solved without cheating. This paradigm allowed experimenters to detect cheating for each unsolvable item. Study 3 examined the validity of the Difference Spotting Task and demonstrated it as a valid tool to assess cheating. The Difference Spotting Task is nonverbal and thus applicable to populations across age, educational level, and culture. In this unsolvable task, participants feel safe in cheating, and experimenters can detect cheating at the item level. The task holds the potential to gain acceptance by many researchers and facilitate the investigation of the underlying processes of cheating behavior.

Keywords Dishonesty · Cheating · Unsolvable paradigm · Difference Spotting Task

Cheating refers to a deliberate action of breaking rules to gain an unfair advantage (Ding et al., 2014; Green, 2004). Cheating behaviors, such as bribes and tax evasion, can cause serious consequences including economic damage. For instance, more than $1 trillion is lost annually because of bribes and tax evasion (Loewen et al., 2013). In efforts to predict and reduce cheating behaviors, behavioral ethics has been devoted to understanding the determinants, boundary conditions, and underlying processes of cheating behavior (Bazerman & Gino, 2012). Several experimental paradigms, such as lottery tasks, ability tests, and unsolvable paradigms, have been developed to measure cheating and address the questions of when, how, and why people cheat (Chou, 2015; Fischbacher & Föllmi-Heusi, 2013; Gneezy, Rockenbach, & Serra-Garcia, 2013; Mazar, Amir, & Ariely, 2008; Zhu et al., 2014). The current study aimed to develop a novel task and assess its ability to measure cheating behaviors.

Lottery is one of the most commonly used paradigms to assess cheating behavior, particularly in the field of economics (e.g., Cohn, Fehr, & Maréchal, 2014; Fischbacher & Föllmi-Heusi, 2013). In this line of studies, participants privately play a lottery using some sort of randomization devices (e.g., rolling a

✉ Hong Li
lihongszu@szu.edu.cn

1 School of Psychology, Shenzhen University, Shenzhen 518060, China

2 Shenzhen Key Laboratory of Affective and Social Cognitive Science, Shenzhen University, Shenzhen 518060, China

3 Research Centre for Brain Function and Psychological Science, Shenzhen University, Shenzhen 518060, China

4 School of Management, Zhejiang University of Technology, Hangzhou 310023, China

5 Institute of Neuromanagement, Zhejiang University of Technology, Hangzhou 310023, China

6 School of Humanities and Management, Guangdong Medical University, Dongguan 523808, China

7 Research Center for Quality of Life and Applied Psychology, Guangdong Medical University, Dongguan 523808, China

8 Department of Psychology, University of Chicago, Chicago, IL, USA

9 Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA

Published online: 10 March 2021

die, flipping a coin, etc.) and receive a payoff determined by their own self-report of the outcome. The randomization device ensures that the true outcome of a single lottery remains unknown except to participants themselves, so that they can disguise the lie at their discretion. In such a setting, although experimenters cannot verify cheating behavior directly, cheating at the group level can be inferred according to a distribution difference between the reported outcome and the expected outcome (Cohn et al., 2014; Fischbacher & Föllmi-Heusi, 2013; Gächter & Schulz, 2016). Extended to repeated rounds of lotteries, this method also allows researchers to infer whether participants cheat at the individual level (e.g., Abe & Greene, 2014; Greene & Paxton, 2009). However, as the inference is based on statistical distributions rather than direct observations of the actual behavior, the exact cheating level is impossible to obtain, both at the individual level and the group level (Gneezy et al., 2013; Moshagen & Hilbig, 2016).

In addition to lottery tasks, another frequently used paradigm is the ability test, especially in psychological research (Mazar et al., 2008; Ruedy, Moore, Gino, & Schweitzer, 2013). Participants are asked to complete a task (e.g., finding two numbers per matrix that added up to 10, unscrambling word jumbles, etc.) and pay themselves according to their own self-report number of correct answers. The answer sheets are shredded by a paper shredder or thrown into a recycle bin or taken back home, so that participants can overreport their performance to earn extra money. To observe cheating at the individual level, experimenters can recover the answer sheets from the recycle bin and compare the actual to the reported performance (Cai, Huang, Wu, & Kou, 2015; Gino & Ariely, 2012; Gino, Schweitzer, Mead, & Ariely, 2011). Nevertheless, participants might be aware of the verifiability of their lies in the recycle bin condition, which could reduce the frequency of cheating compared to the take-home condition (Yaniv & Siniver, 2016).

Taken together, when there is no possibility of getting caught, cheating level can only be inferred from statistical distribution (Fischbacher & Föllmi-Heusi, 2013; Gneezy et al., 2013; Moshagen & Hilbig, 2016). On the other hand, when cheating behavior is observable, such as in the ability test, participants tend to behave honestly in order to not get caught (Yaniv & Siniver, 2016). This conflict between safety to cheat and verifiability of cheating can be resolved in the unsolvable paradigm (Chou, 2015; Evans & Lee, 2011; Karabenick & Srull, 1978; Niiya, Ballantyne, North, & Crocker, 2008; Talwar, Gordon, & Lee, 2007). In this paradigm, participants are asked to complete unsolvable tasks (e.g., unscrambling the unsolvable anagram "DNOEIG") by simply reporting whether they solved each specific item, without providing the exact answer, and then they get paid according to their self-report performance. Consequently, while participants feel safe in cheating because they believe that no one can verify their performance, experimenters can observe the number of unsolvable items on which the participants cheated.

However, classical unsolvable tasks (e.g., unsolvable anagrams) also reveal some shortcomings. First, because they rely on language skills, the task is not applicable to uneducated populations, such as preschool children and illiterate adults. Second, the task is usually filled with solvable items on which participants may also cheat without detection. Third, cheating in these tasks may be confounded with individuals' ability. It may be the case that individuals with lower ability experience more failures and thus have more chances to overreport their performance, showing more cheating behaviors. For instance, in the 20-item matrices task, the number of actually solved matrices varied substantially across individuals (0–20 out of 20 matrices in Cai et al., 2015). Participants with poorer performance in this test had more chances to cheat and thus cheated more ($r = -0.286$, $p < .001$ in Experiment 3, Cai et al., 2015; data can be found online at https://doi.org/10.1016/j.evolhumbehav.2014.09.007).

Thus, this article aimed to develop an improved unsolvable paradigm that overcomes the shortcomings of the lottery tasks, the ability tests, and the unsolvable tasks. First, to avoid the dependence on language skills, Study 1 replaced the anagram task with a new task called the Difference Spotting Task, in which participants were asked to spot one difference between two pictures. As a nonverbal task, the Difference Spotting Task can measure cheating across populations that vary in age, educational level, and culture. Second, to prevent cheating in solvable items and rule out the confounds from individual differences in opportunities to cheat, Study 2 selected very easy items for the solvable items for which cheating would be rare, and the opportunities to overreport performance on all items would be the almost the same across individuals. In the two studies, we selected 10 solvable items and 10 unsolvable items for the mini version of the Difference Spotting Task and 40 solvable items and 40 unsolvable items for the standard version. Finally, to confirm the validity of the Difference Spotting Task, Study 3 additionally used a Die Guessing Task and a Dots Task as measures of cheating behavior.

Study 1: Selection of unsolvable items for the Difference Spotting Task

Methods

Participants Previous research suggests that a sample size larger than 100 is required for reliable item selection (Jones, Smith, & Talley, 2006); consequently, we recruited a large sample of 101 students at Shenzhen University (51 female, mean age = 21.1 years, SD = 1.8). They participated in the study for a bonus based on their performance (¥ 0.2, about

\$0.03, for each solved item). All participants had normal or corrected-to-normal vision and were not color-blind. The reported studies were approved by the Ethics Committee of the School of Psychology, Shenzhen University. Written informed consent was obtained from all participants.

Materials A set of 232 digital color photos (from free internet sources) were used as original stimuli (see Fig. 1a and c for examples). They were randomly assigned to solvable and unsolvable items. The original stimuli were modified by Adobe Photoshop CS3 (Adobe Systems Inc., San Jose, CA, USA). Each solvable item had 10 modifications (see Fig. 1a and b for an example), while each unsolvable item had no modification (see Fig. 1c and d for an example). The modifications included color changing, object copy, and object deletion.

Procedure Participants were seated at a distance of 60 cm. They were asked to play a spot-the-difference game and to find the differences between a pair of similar but different pictures. They were instructed that (1) the number of differences for each pair was greater than or equal to 1; (2) they would earn ¥ 0.2 if they found any difference; (3) they should simply report whether they found any difference (success: "√" or failure: "×"), with no need to indicate the location of the difference on the image; (4) once they spotted one difference, they were asked to make sure that the spotted difference was not a mistake due to lack of attention. The setting offered participants the incentive and opportunity to overreport their performance for extra money and made it seem as if cheating was undetectable. Unbeknownst to the participants, half of the items were unsolvable (see Fig. 1c and d for an example), which made it possible for experimenters to detect cheating in each unsolvable item. The procedure for a single trial of the Difference Spotting Task is depicted in Fig. 2.

Results and discussion

Ninety-five of 101 participants (94.1%) overreported at least one out of 116 unsolvable items. The performance on the unsolvable and solvable items is shown in Fig. 3. Participants reported solving 34.4% (SD = 32.4%) of unsolvable items on average, which is much more than 0% (Wilcoxon signed rank test, z = 8.464, p < .001, n = 101), indicating that overall, participants overreported their performance to earn extra money. Psychometric analysis revealed a high reliability of cheating in the unsolvable items, Cronbach's α = 0.990, and a high average item intercorrelation of 0.462.

Item discrimination and item-total correlation were used as criteria to select unsolvable items of high quality. Participants in the top and bottom 27% were classified as dishonest and honest, respectively. Item discrimination was identified as the difference between the percentage of dishonest participants who cheated on an item and the percentage of honest participants who cheated on that same item. Item discrimination and

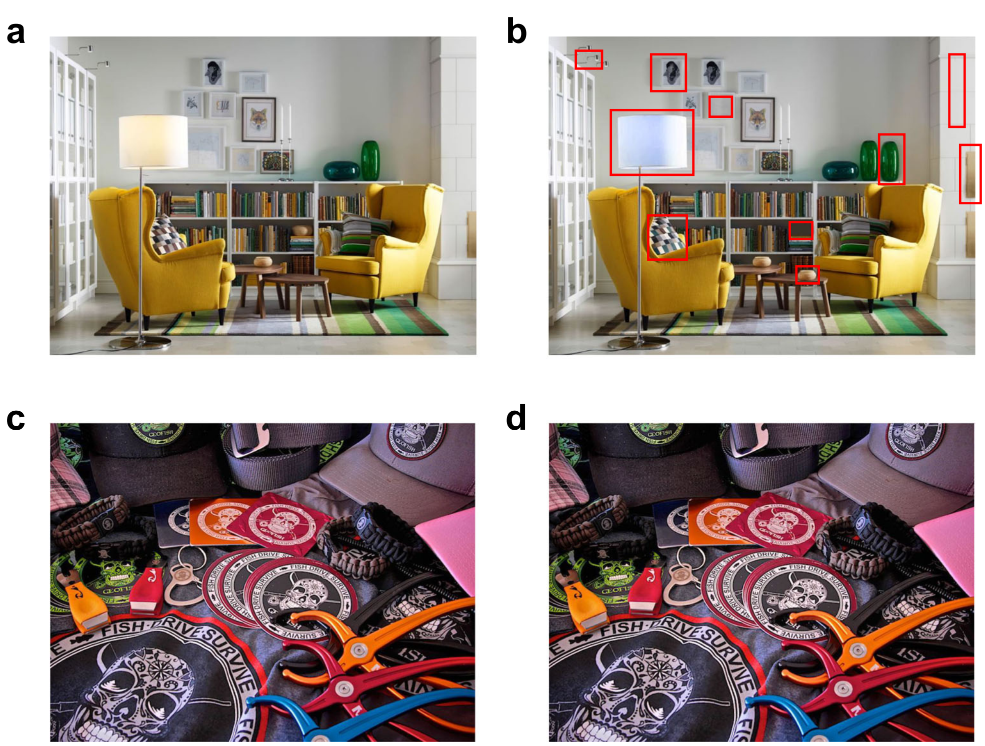

Fig. 1 Examples of the visual stimuli used in Study 1. (**a**) Example of original stimuli in the solvable items. (**b**) Example of modified stimuli in the solvable items. For illustration, the 10 modifications are highlighted by red boxes. (**c**) Example of original stimuli in the unsolvable items. (**d**) Example of unmodified stimuli in the unsolvable items

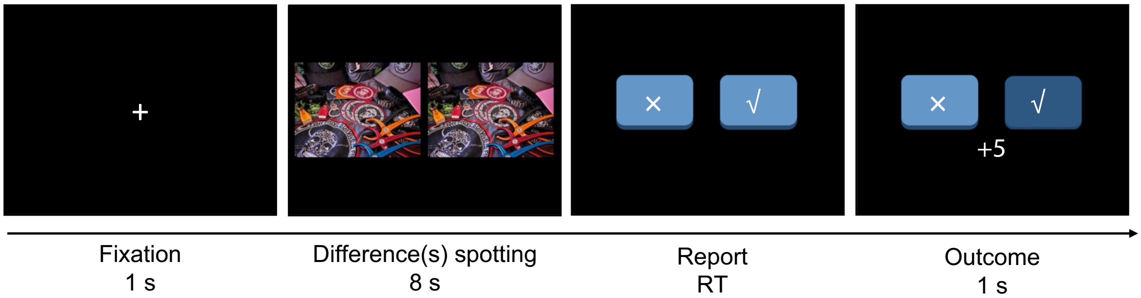

Fig. 2. Sequence of events in a single trial of Difference Spotting Task (in Study 3). Each trial began with a fixation (1 s), followed by a pair of pictures (300 pixels × 400 pixels, 9 cm × 12 cm). After 8 s, the participants were asked to report whether they found any difference, with "√" indicating success and "×" indicating failure (Report screen). Finally, the pressed button turned dark blue and the payoff (¥ 0.5) of the current trial was displayed for 1 s (Outcome screen). This trial shows that the participant cheated in an unsolvable item, in which the paired pictures were identical. *Note:* the payoff for each solved item was ¥ 0.2 in Study 1

item-total correlation were standardized and added together to obtain an index of item quality. As shown in Table S1, items with higher item quality were ranked higher. The top 10 items were used as unsolvable items in the mini version of the Difference Spotting Task, and the top 40 items were used in the standard version of the task. The percentage of solved unsolvable items was 35.1% and 34.9% in the mini and standard versions, respectively. The mini and standard versions showed high reliability, with Cronbach's alphas of 0.944 and 0.982, respectively, and average item intercorrelations of 0.628 and 0.578, respectively. Furthermore, the cheating levels in both the mini and standard versions were highly correlated with those in the 116-item version, with *r*s = 0.963 and 0.989, respectively.

Participants reported solving 91.5% (*SD* = 5.8%) of solvable items on average, which is similar to the percentage of solved items (116 items: *M* = 91.2%, *SD* = 4.9%) when participants had no opportunity to cheat in Study 2 (Mann–Whitney *U* test, *Z* = 0.446, *p* = .655). The increased performance (0.3 out of 116 items on average) could be the result of deliberate cheating or self-deception. Nevertheless, the result indicated that cheating behavior rarely occurred for the solvable items. Notably, as shown in Fig. 3, we found that participants who reported more successes in the unsolvable items also reported more successes in the solvable items, $r = 0.606$, $p < .001$. It may be the case that dishonest individuals performed better in spotting difference(s) than honest individuals. However, to our knowledge, no study has reported an association between visual search ability and dishonesty. It is probable that the ability to spot the difference(s) is independent of dishonesty. Thus, if no one cheated in the solvable items, the self-report performance in the solvable and unsolvable items should be independent. The positive correlation between the performance in the unsolvable and solvable items implied that participants who cheated in the unsolvable items probably also cheated in the solvable items. Considering that the unverifiable number of cheating instances in the solvable items can introduce measurement errors of cheating level, Study 2 aimed to select solvable items, which are easy enough to ensure that there is no need to cheat.

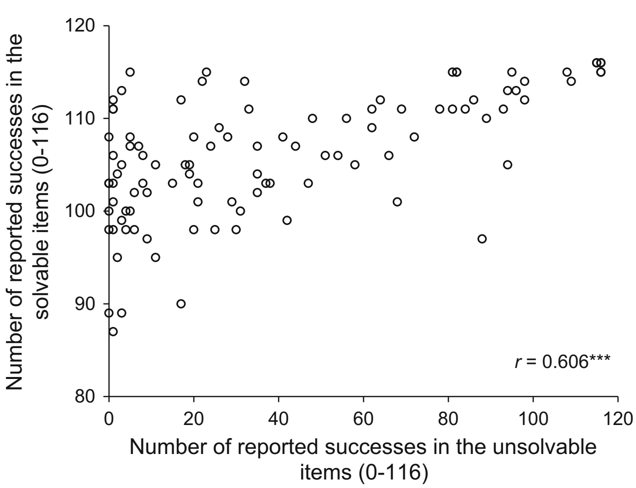

Fig. 3 Reported performance on the unsolvable and solvable items in Study 1 ($N = 101$)

Study 2: Selection of solvable items for the Difference Spotting Task

Methods

Participants Similar to Study 1, we recruited a large sample of 105 students at Shenzhen University (53 female, mean age = 20.2 years, *SD* = 1.6). They participated in the study for a bonus based on their performance (¥ 0.2 for each solved item). All participants had normal or corrected-to-normal vision and were not color-blind.

Materials A set of 125 easy items and 125 difficult items was used. Each easy item contained a pair of pictures with 10 differences, whereas each difficult item contained a pair of pictures with only one difference. For easy items, 116 out of the 125 were identical to the solvable items in Study 1. For difficult items, the left pictures of the 116 difficult items were identical to those of the unsolvable items in Study 1.

Procedure The instructions were identical to Study 1, except that participants had to indicate the identified difference by selecting the area (out of 20 areas) where the difference was located. They would incur a penalty of ¥ 0.2 if the answer was wrong, so they should select failure "×" if they were not certain of their response. This setting did not provide any opportunity to cheat by overreporting performance.

Results and discussion

Participants solved 90.8% of the easy items. Item difficulty and item-total correlation were used as criteria to select solvable items of low discrimination. As shown in Table S2, easier items were ranked higher; for the items that were equally easy, those with smaller absolute item-total correlations were ranked higher. Lower item-total correlation indicated lower correlation with visual search ability. The top 10 easy items were used as solvable items in the mini version of the Difference Spotting Task, and the top 40 items were used in the standard version of the task. For the top 10 items, 100 participants solved all the items, and five participants solved 9 items. For the top 40 items, 53 participants solved all the items, 34 participants solved 39 items, 11 participants solved 38 items, four participant solved 37 items, and three participants solved 35 items, suggesting that in the solvable items, individuals with poor spotting competence might perform as well as those with excellent competence, so that they all had a similar number of opportunities to cheat.

In Study 1 and Study 2, we developed mini and standard versions of the Difference Spotting Task. The results of the mini and standard versions showed a ceiling effect of true performance in the solvable items, which ensures that individuals with high or low visual search abilities would have a similar number of chances to cheat in the task. The results also showed high discrimination of cheating levels in the unsolvable items and well-established reliability of the task, suggesting that the Difference Spotting Task is a useful tool for the assessment of cheating. In Study 3, we aimed to evaluate the validity of the task.

Study 3: Validity of the Difference Spotting Task

We used two additional computerized tasks to evaluate the validity of the Difference Spotting Task. The Die Guessing Task (modified from Greene & Paxton, 2009) is a lottery paradigm in which participants can overreport the accuracy of a predicted die outcome to earn extra money. The Dots Task is a visual task in which participants can intentionally misjudge that the right side of a diagonal line contains more dots than the left side to earn extra money. We chose these two tasks because (1) they are computerized tasks for measuring cheating behavior (e.g., Abe & Greene, 2014; Gino, Norton, & Ariely, 2010; Greene & Paxton, 2009; Mazar & Zhong, 2010; Sharma, Mazar, Alter, & Ariely, 2014) and (2) they consist of multiple-trial tasks in which cheating at the individual level can be inferred.

Methods

Participants Previous research reported medium to large correlations (0.26–0.62) between dishonesty tasks (Gino & Ariely, 2012), so we expected similar correlations and set a correlation of 0.30 in the priori power analysis. The analysis using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) suggested a sample size of 64 to detect a correlation of 0.3 with a one-tailed α level of 0.05 and power of 0.8. Seventy-two students at Shenzhen University (29 female, mean age = 20.8 years, SD = 1.7) participated in the study for a bonus (¥ 0 to ¥ 78) that was based on their performance in the three tasks. All participants had normal or corrected-to-normal vision and were not color-blind. One participant did not complete the Difference Spotting Task, and one participant did not complete the Dots Task. The order of the three tasks was counterbalanced using a Latin square.

Difference Spotting Task Study 3 used the beta version[1] (Tables S3 and S4) of the Difference Spotting Task. The instructions and procedure were identical to those in Study 1. Participants executed four practice items that were all solvable but difficult and received the correct answers after the practice. As we also aimed to evaluate the validity of the mini version, the first 20 trials were restricted to the 10 unsolvable and 10 solvable items in the mini version. The number of self-reported successes in the unsolvable items was used as an index of cheating.

Die Guessing Task The Die Guessing Task was a modified version of the coin-flip task (Abe & Greene, 2014; Greene & Paxton, 2009). Participants were presented with a tossing

[1] In the beta version, the unsolvable items (Table S3) were selected based on an independent small sample in pilot Study 1 (N = 30), and the solvable items (Table S4) were selected based on an independent small sample in pilot Study 2 (N = 30), while in the final version, the items were selected based on a large sample in Study 1 (N = 101) and a large sample in Study 2 (N = 105). For the details of the small samples, see Supplementary Material. As shown in Table S1, S2, S3, and S4, 18 unsolvable items were included both in the beta version and the final version; 30 solvable items were included both in the beta version and the final version. The cheating levels in the beta and final version were highly correlated in Study 1 (101 participants), r = 0.985, p <.001.

die video and were asked to predict the outcome of the die (1, 2, and 3 were classified as "low", while 4, 5, and 6 were classified as "high"). We told participants that (1) the financial rewards for accuracy were varied (¥ 0, ¥ 0.1, or ¥ 0.5) to test whether people made more accurate predictions when they were motivated to predict accurately; and (2) to test the idea that people's ability to predict the future would be disrupted if they had to record their predictions externally, participants should sometimes report their predictions in advance, but in most cases they should keep their predictions private and simply tell us whether or not their predictions were correct. The settings gave them the opportunity and incentive to cheat. Experimenters could draw inferences on a distribution difference between the reported outcome and the expected outcome.

The task consisted of 90 trials. Each financial reward (¥ 0, ¥ 0.1, or ¥ 0.5) had 30 trials: 6 trials of public predictions and 24 trials of private predictions. Each trial began with a fixation (0.5 s), followed by a reward cue (¥ 0, ¥ 0.1, or ¥ 0.5) for accuracy (2 s) and a tossing die video (approximately 4 s). Then participants were asked to privately predict the outcome of the die and record this prediction by clicking on one of two buttons (i.e., "high" or "low", public condition) or clicking the buttons (i.e., "※" or "※", private condition) randomly. After that, they observed the outcome of the die (approximately 2 s) and were asked to indicate whether their predictions were accurate. Finally, they saw the payoff (¥ 0, ¥ 0.1, or ¥ 0.5) for 1 s. The number of self-reported wins in the 24 private trials with a reward of ¥ 0.5 was used as an index of cheating behavior.

Dots Task In the Dots Task (adapted from Mazar & Zhong, 2010; see also Gino, Norton, & Ariely, 2010; Sharma, Mazar, Alter, & Ariely, 2014), we presented participants with a square of 20 dots divided into two by a diagonal line for 1 s and asked them to identify which side of the diagonal (left or right) contained more dots by pressing the corresponding key ("F" or "J"). They were instructed that because most people find it easier to estimate the number of dots on the left rather than right side of the line, they would earn ¥ 0.01 for each press on the left key which indicated more dots on the left side, and ¥ 0.1 for each press on the right key which indicated more dots on the right side. They were also asked to be as accurate as possible because their results would be very important for the study. Therefore, they experienced a conflict between honesty and selfish gain when they perceived more dots on the left.

In the first phase, participants performed 100 practice trials for which they received feedback about their trial-by-trial and cumulative earnings with hypothetical money. In the second phase, they performed 200 trials for which they received feedback with real money. Each set of 100 trials consisted of 35 trials in which there were clearly more dots on the left (i.e., 14 or 15 dots on the left), 15 trials in which there were clearly more dots on the right (i.e., 14 or 15 dots on the right), 25 trials in which there were ambiguously more dots on the left (i.e., 11 dots on the left), and 25 trials in which there were ambiguously more dots on the right (i.e., 11 dots on the right). The number of times that participants judged more dots on the right in 50 ambiguous trials where there were actually more dots on left was used as an index of cheating.

Results and discussion

Descriptive statistics and correlations for measures of cheating are shown in Table 1.

Participants lied to earn extra money in all three tasks at the group level. In the Difference Spotting Task, 53 out of 71 participants (74.6%) overreported their performance on the unsolvable items. They reported that they solved 17.2% (SD =23.9%) of unsolvable items on average, which was significantly more than 0% (Wilcoxon signed rank test, z = 6.343, p < .001, n = 71). Likewise, in the Die Guessing Task, the average accuracy in the private trials with a reward of ¥ 0.5 (M = 63.6%, SD = 15.6%) was significantly higher than the full honesty benchmark of 50% (Wilcoxon signed rank test, z = 5.662, p < .001, n = 72). Similarly, in the Dots Task, participants performed better in trials where there were ambiguously more dots on the right (M = 86.5%, SD = 9.7%) than in trials where there were ambiguously more dots on the left (M = 79.3%, SD = 15.4%; Wilcoxon signed rank test, z = 2.181, p = .029, n = 71).

The Difference Spotting Task again exhibited robust psychometric properties. As shown in Table 1, psychometric analysis revealed high reliability of cheating levels in the Difference Spotting Task, with Cronbach's alphas of 0.793 and 0.961, and average item intercorrelations of 0.279 and 0.386 for the mini and standard versions, respectively. Furthermore, correlation analysis supported the validity of the Difference Spotting Task as a behavioral measure of cheating. Cheating levels in the Difference Spotting Task were moderately correlated with cheating levels in the Die Guessing task (r = 0.515) and the Dots Task (r = 0.393). Notably, participants solved 97.2% (SD = 6.6%) and 95.7% (SD = 5.1%) of solvable items in the mini and standard versions, respectively. The results showed a ceiling effect of performance and very low incidence of cheating on the solvable items, suggesting that the numbers of chances for each participant to cheat were almost equal. The negligibility of cheating in solvable items was further supported by the low correlations between the performance in the solvable items and the cheating levels in the three cheating tasks, with a range of rs = [−0.073, 0.058] and [−0.059, 0.119] for the mini and standard versions, respectively.

Results also suggest that the psychometric indicators of the Difference Spotting Task are at least as good as those of the other two behavioral measures, or even better. We conducted

Table 1 Descriptive statistics and correlations for measures of cheating in Study 3

Task	Mean time (s)	No. of trials	α	Average item intercorrelation	Range	Mean	SD	Correlations		
								1	2	3
1. Cheating level in standard version of Difference Spotting Task	887	40	0.961	0.386	0–39	6.9	9.6			
2. Cheating level in mini version of Difference Spotting Task	227	10	0.793	0.279	0–9	1.6	2.2	.903***		
3. Cheating level in Die Guessing Task	1169	24	0.634	0.067	8–23	15.3	3.7	.515***	.463***	
4. Cheating level in ambiguous more-left trials (Dots Task)	367	50	0.883	0.142	0–32	10.3	7.7	.393**	.299*	.295*

*$p < .05$; **$p < .01$; ***$p < .001$

the comparison of Cronbach's alphas with the R package cocron (Diedenhofen & Musch, 2016) and found that the Cronbach's alpha of the Difference Spotting Task (α = 0.961) was higher than that of the Die Guessing Task (0.634) and Dots Task (0.883), χ^2 (1) = 83.7, $p < .001$; χ^2 (1) = 22.0, $p < .001$, respectively. Considering that Cronbach's alpha increases as the number of items increases, the comparison of Cronbach's alphas would be unfair for instruments with a small number of items. Therefore, we used item intercorrelation, a straightforward indicator of internal consistency that is independent of item number (Clark & Watson, 1995). Likewise, we found that the item intercorrelations of the Difference Spotting Task (M±SD: rs = 0.386 ±0.132; Fisher's zs = 0.417±0.160) were higher than those of the Die Guessing Task (0.067±0.126; 0.068±0.129) and Dots Task (0.142±0.143; 0.146±0.149), $t(1054) = 32.697$, $p < .001$; $t(2003) = 38.639$, $p < .001$, respectively. Additionally, correlations between the cheating on the Difference Spotting Task and on the other two behavioral measures were higher than the correlation between the other two behavioral measures. However, the comparison of correlation coefficients using the computational tool cocor (Diedenhofen & Musch, 2015; Zou, 2007) showed that the differences did not reach significance (0.515 vs. 0.295, 95% Zou's confidence interval is [−0.009, 0.450]; 0.393 vs. 0.295, 95% Zou's confidence interval is [−0.117, 0.313]).

General discussion

This project introduced a new task to measure cheating behavior at the item level: the Difference Spotting Task. Across three studies, we established its reliability and validity and demonstrated how this novel task setting can solve issues of existing assessments of cheating.

First, the task allowed participants to feel safe in cheating without being caught. Studies 1 and 3 showed that, under a context with no fear of being caught or punished, a large portion of participants cheated at least once (94.1% in Study 1 and 74.6% in Study 3). These proportions are comparable to the 82% of participants who were not afraid of being caught cheating in another study (e.g., at home; Yaniv & Siniver, 2016), supporting the proposal that the Difference Spotting Task has the advantage of unrestricted cheating opportunity. Participants' belief that it was safe to cheat is an important issue considering that the fear of getting caught decreases the frequency of cheating, which may cause a floor effect and render the experimental manipulation insensitive to cheating. For instance, a meta-analysis revealed that dishonesty behaviors substantially decreased in the matrix task in which dishonesty could be verified individually (Gerlach, Teodorescu, & Hertwig, 2019). Another study showed that cheating levels were very low in a task where cheating could be detected. The authors failed to detect an effect of cognitive control impairment on cheating behavior. However, when they used another behavioral task to increase the level of cheating, they successfully obtained a significant effect (Pitesa, Thau, & Pillutla, 2013). Hence, these studies corroborate the notion that behavioral tasks in which participants feel that it is safe to cheat, such as the Difference Spotting Task, are ideal for measuring cheating and detecting differences across conditions.

Second, experimenters are able to verify cheating for each unsolvable item, whereas the commonly used paradigms only allow for indirect inference of cheating. For example, lottery tasks usually use self-report payoffs as a proxy for actual dishonesty (Cohn et al., 2014; Fischbacher & Föllmi-Heusi, 2013; Gächter & Schulz, 2016). However, this method adds random noise to the results and underestimates the effects of experimental manipulation on dishonesty (Moshagen & Hilbig, 2016). Indeed, a meta-analysis found no effect of material incentives on cheating in lottery tasks, but did find a significant effect in another task (Gerlach et al., 2019). Moreover, random noise reduces statistical power and increases the sample size required (Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007). The Difference Spotting Task can address these issues, since it enables experimenters to directly observe rather than indirectly infer cheating. In this way, our task outperforms other tasks such as the Die

Guessing Task and Dots Task in terms of internal reliability and convergent validity, as shown in Table 1. Furthermore, because of the improved psychometric properties, the impact of score unreliability on the insensitivity of effect size is reduced, so that the sample size required for achieving the given statistical power is smaller in this task than in lottery tasks.

Third, this task also allows experimenters to record the reaction time for each item. The measure of reaction time in our task is able to address the disputes over the cognitive mechanisms of cheating behavior, such as the competing "Will" and "Grace" hypotheses. Shalvi et al. (2012) found that time pressure increased the frequency of cheating, suggesting that cheating is an automatic response of the self-serving tendency towards financial interests. This interpretation is consistent with the "Will" hypothesis that understands honest decision as the result of active resistance to temptation. On the other hand, the "Grace" hypothesis regards honest behavior as resulting from the absence of temptation. Greene and Paxton (2009) compared neural activity in honest and dishonest groups. They found that, relative to a control condition with no opportunity for dishonest gain, the honest group showed no additional activity in the cognitive control network for honest decisions, while the dishonest group exhibited activity in the cognitive control network for both honest and dishonest decisions (Greene & Paxton, 2009), supporting the view that honesty reflects the absence of temptation. To address the debate over automatic or controlled processes, future studies can compare the reaction times between the honest and dishonest trials for honest and dishonest individuals by using the Difference Spotting Task, given that this task allows experimenters to conduct reaction time and trial-by-trial analysis.

Last but not least, the task is easily applicable for research. In contrast with other tasks, the Difference Spotting Task does not rely on language skills, so its use can be extended and generalized to any population with normal vision. Unlike unsolvable anagrams (Chou, 2015; Karabenick & Srull, 1978; Niiya et al., 2008), our picture-based task allows researchers to compare cheating behaviors across different populations, varying in age, educational level, and culture. The task is simple and intuitive, and in most cases, participants can fully understand the instruction without further clarification from the experimenter. This reduces potential external variables coming from the experimenter, such as nonverbal cues. We have disclosed the instructions, stimuli, and codes for the task to further contribute to easy accessibility for researchers (https://osf.io/s8qdk/). Computer and web versions of the task are recommended, but a paper-pencil version is also possible with a timer and color-printed pictures. In addition, to shorten the length of administration time, we also developed a mini version of the task and provided data supporting its reliability and validity.

Nevertheless, some limitations should be noted. First, although the task instructed participants to look carefully, they may also have falsely believed that they saw a difference in the unsolvable items. The task cannot exclude this possibility of self-deception in the overreporting. Study 2 found that participants self-reported successes but pointed out the wrong locations of the differences in 1.2% of difficult items when participants had no opportunity to cheat. These errors may have resulted from self-deception, false memory of the location number, typing error, etc. The low error rate suggests a low incidence of self-deception. Second, to make participants feel that it was safe to cheat, we deceived them about the solvability of the Difference Spotting Task. This important ethical issue constitutes a disadvantage of the task as compared to lottery tasks, especially in economic studies (Hertwig & Ortmann, 2001). To avoid deceiving participants, it is possible to replace the unsolvable items with very difficult but solvable items (e.g., Hoffmann, Diedenhofen, Verschuere, & Musch, 2015). Therefore, we created a non-deception mini version of the task, in which we selected a list of practical unsolvable items from the difficult items in Study 2 (Table S5). The high correlation ($r = 0.923$) between this version and the standard version in Study 1 suggests its validity. Notably, in this version of the task, we cannot exclude the possibility that participants may actually solve some of the very difficult items, which can bias the measure of cheating.

To conclude, we developed a new task called the Difference Spotting Task that measures cheating behavior, and provided empirical data supporting its high reliability, validity, and ease of use. This novel task holds the potential to gain acceptance by many researchers and to improve the understanding of the underlying processes of cheating behavior.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-020-01526-w.

Acknowledgments We thank Miss Haiqi Lin, Miss Yawen Mao, Mr. Xiaoguang Yang, and Miss Liying Wang for their assistance in data collection.

Author Contributions J. L., Q. S., J. Z., and H. L. designed the experiments. J. L. performed the experiments. J. L. and J. Z. analyzed the data. J. L., Q. S., U. B., W. C., J. D., and H. L. wrote the manuscript.

Funding This work was supported by grants from the National Natural Science Foundation of China (31600928) and Shenzhen Basic Research Project (JCYJ20190808154209286) to Jinting Liu, and Guangdong Key Project (2018B030335001) to Hong Li.

Compliance with ethical standards

Declaration of Conflicting Interests The authors declare that they have no conflict of interest with respect to their authorship or the publication of this article.

References

Abe, N., & Greene, J. D. (2014). Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *The Journal of Neuroscience*, *34*(32), 10564–10572. https://doi.org/10.1523/JNEUROSCI.0217-14.2014

Bazerman, M. H., & Gino, F. (2012). Behavioral ethics: Toward a deeper understanding of moral judgment and dishonesty. *Annual Review of Law and Social Science*, *8*(1), 85–104. https://doi.org/10.1146/annurev-lawsocsci-102811-173815

Cai, W., Huang, X., Wu, S., & Kou, Y. (2015). Dishonest behavior is not affected by an image of watching eyes. *Evolution and Human Behavior*, *36*(2), 110–116. https://doi.org/10.1016/j.evolhumbehav.2014.09.007

Chou, E. Y. (2015). What's in a name? The toll e-signatures take on individual honesty. *Journal of Experimental Social Psychology*, *61*, 84–95. https://doi.org/10.1016/j.jesp.2015.07.010

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, *516*(7529), 86–89. https://doi.org/10.1038/nature13977

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach ' s alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

Ding, X. P., Omrin, D. S., Evans, A. D., Fu, G., Chen, G., & Lee, K. (2014). Elementary school children's cheating behavior and its cognitive correlates. *Journal of Experimental Child Psychology*, *121*(1), 85–95. https://doi.org/10.1016/j.jecp.2013.12.005

Evans, A. D., & Lee, K. (2011). Verbal deception from late childhood to middle adolescence and its relation to executive functioning skills. *Developmental Psychology*, *47*(4), 1108–1116. https://doi.org/10.1037/a0023425

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547. https://doi.org/10.1111/jeea.12014

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496–499. https://doi.org/10.1038/nature17160

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1–44. https://doi.org/10.1037/bul0000174

Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*(3), 445–459. https://doi.org/10.1037/a0026406

Gino, F., Norton, M. I., & Ariely, D. (2010). The counterfeit self: The deceptive costs of faking it. *Psychological Science*, *21*(5), 712–720. https://doi.org/10.1177/0956797610366545

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 191–203. https://doi.org/10.1016/j.obhdp.2011.03.001

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, *93*, 293–300. https://doi.org/10.1016/j.jebo.2013.03.025

Green, S. P. (2004). Cheating. *Law and Philosophy*, *23*, 137–185.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, *106*(30), 12506–12511. https://doi.org/10.1073/pnas.0900152106

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403. https://doi.org/10.1017/s0140525x01004149

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*, *62*(6), 403–414. https://doi.org/10.1027/1618-3169/a000304

Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of Test Development* (pp. 487–525). New York: Routledge. https://doi.org/10.4324/9780203874776.ch22

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 81–90. https://doi.org/10.22237/jmasm/1177992480

Karabenick, S. A., & Srull, T. K. (1978). Effects of personality and situational variation in locus of control on cheating: determinants of the "congruence effect". *Journal of Personality*, *46*(1), 72–95. https://doi.org/10.1111/j.1467-6494.1978.tb00603.x

Loewen, P. J., Dawes, C. T., Mazar, N., Johannesson, M., Koellinger, P., & Magnusson, P. K. E. (2013). The heritability of moral standards for everyday dishonesty. *Journal of Economic Behavior & Organization*, *93*, 363–366. https://doi.org/10.1016/j.jebo.2013.05.001

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Mazar, N., & Zhong, C.-B. (2010). Do green products make us better people? *Psychological Science*, *21*(4), 494–498. https://doi.org/10.1177/0956797610363538

Moshagen, M., & Hilbig, B. E. (2016). The statistical analysis of cheating paradigms. *Behavior Research Methods*. https://doi.org/10.3758/s13428-016-0729-x

Niiya, Y., Ballantyne, R., North, M. S., & Crocker, J. (2008). Gender, contingencies of self-worth, and achievement goals as predictors of academic cheating in a controlled laboratory setting. *Basic and Applied Social Psychology*, *30*(1), 76–83. https://doi.org/10.1080/01973530701866656

Pitesa, M., Thau, S., & Pillutla, M. M. (2013). Cognitive control and socially desirable behavior: The role of interpersonal impact. *Organizational Behavior and Human Decision Processes*, *122*(2), 232–243. https://doi.org/10.1016/j.obhdp.2013.08.003

Ruedy, N. E., Moore, C., Gino, F., & Schweitzer, M. E. (2013). The cheater's high: The unexpected affective benefits of unethical behavior. *Journal of Personality and Social Psychology*, *105*(4), 531–548. https://doi.org/10.1037/a0034231

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science*, *23*(10), 1264–1270. https://doi.org/10.1177/0956797612443835

Sharma, E., Mazar, N., Alter, A. L., & Ariely, D. (2014). Financial deprivation selectively shifts moral standards and compromises moral decisions. *Organizational Behavior and Human Decision Processes*, *123*(2), 90–100. https://doi.org/10.1016/j.obhdp.2013.09.001

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810. https://doi.org/10.1037/0012-1649.43.3.804

Yaniv, G., & Siniver, E. (2016). The (honest) truth about rational dishonesty. *Journal of Economic Psychology*, *53*, 131–140. https://doi.org/10.1016/j.joep.2016.01.002

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., … Hsu, M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, *17*(10), 1319–1321. https://doi.org/10.1038/nn.3798

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

Open practices The instructions, stimuli, and codes for the Difference Spotting Task, as well as all the data are available through the Open Science Framework repository: https://doi.org/https://osf.io/s8qdk/. The large samples of Study 1 (https://doi.org/https://aspredicted.org/qj7bj.pdf) and Study 2 (https://doi.org/https://aspredicted.org/jh6jp.pdf) were preregistered.