# A costly Bayesian implementable social choice function may not be truthfully implementable

Wu, Haoyang

11 September 2016

# A costly Bayesian implementable social choice function may not be truthfully implementable

Haoyang Wu *

*Wan-Dou-Miao Research Lab, Room 301, Building 3, 718 WuYi Road, Shanghai, 200051, China.*

**Abstract**

The revelation principle is a fundamental theorem in many economics fields. In this paper, we construct a simple labor model to show that a social choice function which can be implemented costly in Bayesian Nash equilibrium may not be truthfully implementable. The key point is the strategy cost property given in Section 4: In the direct mechanism, each agent only reports a type and will not pay the strategy cost which would be paid by himself when playing strategies in the original indirect mechanism. As a result, the revelation principle may not hold when agents' strategies are costly in the indirect mechanism.

JEL codes: D71, D82

*Key words:* Revelation principle; Game theory; Mechanism design; Auction theory.

## 1 Introduction

The revelation principle plays an important role in microeconomics theory and has been applied to many other fields such as auction theory, mechanism design *etc.* According to the wide-spread textbook given by Mas-Colell, Whinston and Green (Page 884, Line 24 [1]): "*The implication of the revelation principle is ... to identify the set of implementable social choice functions in Bayesian Nash equilibrium, we need only identify those that are truthfully implementable.*" Related definitions about the revelation principle can be seen in Appendix, which are cited from Section 23.B and 23.D of MWG's textbook[1].

* Corresponding author.
*Email address:* 18621753457@163.com, Tel: 86-18621753457 (Haoyang Wu).

Generally speaking, some costs are required for a social choice function to be implemented by a mechanism. There are two different kinds of costs possibly occurred in a mechanism: 1) *strategy costs*, which are possibly occurred when agents play strategies; and 2) *misreporting costs*, which are possibly occurred when agents report types falsely. [1] In the traditional literature of mechanism design, costs are usually referred to the former. Recently, some researchers began to investigate misreporting costs. For every type $\theta$ and every type $\hat{\theta}$ an agent might misreport, Kephart and Conitzer [2] defined a cost function as $c(\theta, \hat{\theta})$ for doing so. Traditional mechanism design is just the case where $c(\theta, \hat{\theta}) = 0$ everywhere, and partial verification is a special case where $c(\theta, \hat{\theta}) \in \{0, \infty\}$ [3–5]. Kephart and Conitzer [2] proposed that when reporting truthfully is costless and misreporting can be costly, the revelation principle can fail to hold.

Despite these accomplishments, so far people seldom consider the two different kinds of costs simultaneously. The aim of this paper is to investigate the justification of revelation principle when both of two kinds of costs are considered. The paper is organized as follows. In Section 2, we construct a labor model, then define a social choice function $f$ and an indirect mechanism, in which agents' strategies are costly. In Section 3, we prove $f$ can be implemented by the indirect mechanism in Bayesian Nash equilibrium. In Section 4, we propose a strategy cost property and point out that the revelation principle may not hold even if misreporting costs are zero and only strategy costs occur. In Section 5, we prove that $f$ is not truthfully implementable in Bayesian Nash equilibrium, which contradicts the revelation principle. Finally, Section 6 draws conclusions.

## 2 A labor model

Here we construct a labor model which uses some ideas from the first-price sealed auction model in Example 23.B.5 [1] and the signaling model [1,6]. There are one firm and two workers. Worker 1 and Worker 2 differ in the number of units of output they produce if hired by the firm, which is denoted by productivity type. The firm wants to hire a worker with productivity as high as possible, and the two workers compete for this job offer.

For simplicity, we make the following assumptions:
1) The possible productivity types of two workers are: $\theta_L$ and $\theta_H$, where $\theta_H > \theta_L > 0$. Each worker $i$'s productivity type $\theta_i$ $(i = 1, 2)$ is his private information.
2) There is a certification that the firm can announce as a hire criterion. The

[1] It is usually assumed that each agent can report his true type with zero cost.

education level corresponding to the certification is $e_H > 0$. Each worker decides by himself whether to get the certification or not, hence the possible education levels are $e_H$ and 0. Each worker $i$'s education level $e_i$ $(i = 1, 2)$ is observable to the firm. Education does nothing for a worker's productivity.
3) The strategy cost of obtaining education $e_i$ for a worker $i$ $(i = 1, 2)$ of productivity type $\theta_i$ is given by a function $c(e_i, \theta_i) = e_i/\theta_i$. That is, the strategy cost is lower for a high-productivity worker.
4) The misreporting cost for a low-productivity worker to report the high productivity type $\theta_H$ is a fixed value $c' > 0$. In addition, a high-productivity worker is assumed to report the low productivity type $\theta_L$ with zero cost.

The labor model's outcome is represented by a vector $(y_1, y_2)$, where $y_i$ denotes the probability that worker $i$ gets the job offer with wage $w > 0$ which is chosen by the firm. Recall that the firm does not know the exact productivity types of two workers, but its aim is to hire a worker with productivity as high as possible. This aim can be represented by a social choice function $f(\theta) = (y_1(\theta), y_2(\theta))$, in which $\theta = (\theta_1, \theta_2)$,

$$y_1(\theta) = \begin{cases} 1, & \text{if } \theta_1 > \theta_2 \\ 0.5, & \text{if } \theta_1 = \theta_2 \\ 0, & \text{if } \theta_1 < \theta_2 \end{cases}, \; y_2(\theta) = \begin{cases} 1, & \text{if } \theta_1 < \theta_2 \\ 0.5, & \text{if } \theta_1 = \theta_2 \\ 0, & \text{if } \theta_1 > \theta_2 \end{cases} \tag{1}$$

In order to implement the above $f(\theta)$, the firm designs an indirect mechanism $\Gamma = (S_1, S_2, g)$ as follows: For each worker $i = 1, 2$, conditional on his type $\theta_i \in \{\theta_L, \theta_H\}$, he chooses the education level as a bid $b_i : \{\theta_L, \theta_H\} \to \{0, e_H\}$. The strategy set $S_i$ is the set of all possible bids, and the outcome function $g$ is defined as:

$$g(b_1, b_2) = (p_1, p_2) = \begin{cases} (1, 0), & \text{if } b_1 > b_2 \\ (0.5, 0.5), & \text{if } b_1 = b_2 \\ (0, 1), & \text{if } b_1 < b_2 \end{cases} \tag{2}$$

where $p_i$ $(i = 1, 2)$ is the probability that worker $i$ gets the job offer.

Let $u_0$ be the utility of the firm, and $u_1, u_2$ be the utilities of worker $1, 2$ in the indirect mechanism $\Gamma$ respectively, then $u_0(b_1, b_2) = p_1\theta_1 + p_2\theta_2 - w$, and for $i, j = 1, 2$, $i \neq j$,

$$u_i(b_i, b_j; \theta_i) = \begin{cases} w - b_i/\theta_i, & \text{if } b_i > b_j \\ 0.5w - b_i/\theta_i, & \text{if } b_i = b_j \\ -b_i/\theta_i, & \text{if } b_i < b_j \end{cases} \tag{3}$$

The item "$b_i/\theta_i$" occurred in Eq (3) is just the strategy cost paid by agent $i$ of type $\theta_i$ when he performs the strategy $b_i(\theta_i)$ in the indirect mechanism.

Suppose the conservative utilities of worker 1 and worker 2 are both zero, then the individual rationality (IR) constraints are: $u_i(b_i, b_j; \theta_i) \geq 0$, $i = 1, 2$.

## 3 $f$ is Bayesian implementable

**Proposition 1:** If $w \in (2e_H/\theta_H, 2e_H/\theta_L)$, the social choice function $f(\theta)$ given in Eq (1) can be implemented by the indirect mechanism $\Gamma$ in Bayesian Nash equilibrium.
**Proof:** Consider a separating strategy, *i.e.*, workers with different productivity types choose different education levels,

$$b_1(\theta_1) = \begin{cases} e_H, & \text{if } \theta_1 = \theta_H \\ 0, & \text{if } \theta_1 = \theta_L \end{cases}, \; b_2(\theta_2) = \begin{cases} e_H, & \text{if } \theta_2 = \theta_H \\ 0, & \text{if } \theta_2 = \theta_L \end{cases}. \tag{4}$$

Now let us check whether this separating strategy yields a Bayesian Nash equilibrium. Assume $b_j^*(\theta_j)$ takes this form, *i.e.*,

$$b_j^*(\theta_j) = \begin{cases} e_H, & \text{if } \theta_j = \theta_H \\ 0, & \text{if } \theta_j = \theta_L \end{cases}, \tag{5}$$

then we consider worker $i$'s problem ($i \neq j$). For each $\theta_i \in \{\theta_L, \theta_H\}$, worker $i$ solves a maximization problem: $\max_{b_i} h(b_i, \theta_i)$, where by Eq (3) the object function is

$$h(b_i, \theta_i) = (w - b_i/\theta_i)P(b_i > b_j^*(\theta_j)) + (0.5w - b_i/\theta_i)P(b_i = b_j^*(\theta_j)) - (b_i/\theta_i)P(b_i < b_j^*(\theta_j)) \tag{6}$$

We discuss this maximization problem in four different cases:
1) Suppose $\theta_i = \theta_j = \theta_L$, then $b_j^*(\theta_j) = 0$ by Eq (5).

$$\begin{aligned} h(b_i, \theta_i) &= (w - b_i/\theta_L)P(b_i > 0) + (0.5w - b_i/\theta_L)P(b_i = 0) - (b_i/\theta_L)P(b_i < 0) \\ &= \begin{cases} w - e_H/\theta_L, & \text{if } b_i = e_H \\ 0.5w, & \text{if } b_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w < 2e_H/\theta_L$, then $h(e_H, \theta_i) < h(0, \theta_i)$, which means the optimal value of $b_i(\theta_i)$ is 0. In this case, $b_i^*(\theta_L) = 0$.

2) Suppose $\theta_i = \theta_L$, $\theta_j = \theta_H$, then $b_j^*(\theta_j) = e_H$ by Eq (5).

$$\begin{aligned} h(b_i, \theta_i) &= (w - b_i/\theta_L)P(b_i > e_H) + (0.5w - b_i/\theta_L)P(b_i = e_H) - (b_i/\theta_L)P(b_i < e_H) \\ &= \begin{cases} 0.5w - e_H/\theta_L, & \text{if } b_i = e_H \\ 0, & \text{if } b_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w < 2e_H/\theta_L$, then $h(e_H, \theta_i) < h(0, \theta_i)$, which means the optimal value of $b_i(\theta_i)$ is 0. In this case, $b_i^*(\theta_L) = 0$.

3) Suppose $\theta_i = \theta_H$, $\theta_j = \theta_L$, then $b_j^*(\theta_j) = 0$ by Eq (5).

$$\begin{aligned} h(b_i, \theta_i) &= (w - b_i/\theta_H)P(b_i > 0) + (0.5w - b_i/\theta_H)P(b_i = 0) - (b_i/\theta_H)P(b_i < 0) \\ &= \begin{cases} w - e_H/\theta_H, & \text{if } b_i = e_H \\ 0.5w, & \text{if } b_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w > 2e_H/\theta_H$, then $h(e_H, \theta_i) > h(0, \theta_i)$, which means the optimal value of $b_i(\theta_i)$ is $e_H$. In this case, $b_i^*(\theta_H) = e_H$.

4) Suppose $\theta_i = \theta_j = \theta_H$, then $b_j^*(\theta_j) = e_H$ by Eq (5).

$$\begin{aligned} h(b_i, \theta_i) &= (w - b_i/\theta_H)P(b_i > e_H) + (0.5w - b_i/\theta_H)P(b_i = e_H) - (b_i/\theta_H)P(b_i < e_H) \\ &= \begin{cases} 0.5w - e_H/\theta_H, & \text{if } b_i = e_H \\ 0, & \text{if } b_i = 0 \end{cases}. \end{aligned}$$

Thus, if $w > 2e_H/\theta_H$, then $h(e_H, \theta_i) > h(0, \theta_i)$, which means the optimal value of $b_i(\theta_i)$ is $e_H$. In this case, $b_i^*(\theta_H) = e_H$.

From the above four cases, it can be seen that if the wage $w \in (2e_H/\theta_H, 2e_H/\theta_L)$, the strategy $b_i^*(\theta_i)$ of worker $i$

$$b_i^*(\theta_i) = \begin{cases} e_H, & \text{if } \theta_i = \theta_H \\ 0, & \text{if } \theta_i = \theta_L \end{cases} \tag{7}$$

is the optimal response to the strategy $b_j^*(\theta_j)$ of worker $j$ ($j \neq i$) given in Eq (5). Therefore, the strategy profile $(b_1^*(\theta_1), b_2^*(\theta_2))$ is a Bayesian Nash equilibrium of the game induced by $\Gamma$.

Now let us investigate whether the wage $w \in (2e_H/\theta_H, 2e_H/\theta_L)$ satisfies the individual rationality (IR) constraints. Following Eq (3) and Eq (7), the (IR) constraints are changed into: $0.5w - b_H/\theta_H > 0$. Obviously, $w \in (2e_H/\theta_H, 2e_H/\theta_L)$ satisfies the (IR) constraints.

In summary, if $w \in (2e_H/\theta_H, 2e_H/\theta_L)$, then by Eq(2) and Eq(7), for any $\theta = (\theta_1, \theta_2)$, where $\theta_1, \theta_2 \in \{\theta_L, \theta_H\}$, there holds:

$$g(b_1^*(\theta_1), b_2^*(\theta_2)) = \begin{cases} (1, 0), & \text{if } \theta_1 > \theta_2 \\ (0.5, 0.5), & \text{if } \theta_1 = \theta_2 \\ (0, 1), & \text{if } \theta_1 < \theta_2 \end{cases}, \tag{8}$$

which is just the social choice function $f(\theta)$ given in Eq (1). □

## 4 Strategy cost property

Before discussing the truthful implementation problem of a costly Bayesian implementable social choice function, we first cite the basic idea behind the revelation principle given in MWG's textbook (Page 884, Line 16, [1]): "If in mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$, each agent finds that, when his type is $\theta_i$, choosing $s_i^*(\theta_i)$ is his best response to the other agents' strategies, then if we introduce mediator who says '*Tell me your type*, $\theta_i$, *and I will play* $s_i^*(\theta_i)$ *for you*', each agent will find truth telling to be an optimal strategy given that all other agents tell the truth. That is, truth telling will be a Bayesian Nash equilibrium of this direct revelation game".

Although this basic idea looks reasonable, we propose that behind the mediator's announcement "*Tell me your type*, $\theta_i$, *I will play* $s_i^*(\theta_i)$ *for you*", an assumption should be added to make this announcement *credible*: after receiving each agent $i$'s report $\theta_i$ ($i = 1, \cdots, I$), in order to be able to play $s_i^*(\theta_i)$ for agent $i$, the mediator should also pay the strategy cost which would be paid by agent $i$ himself when carrying out $s_i^*(\theta_i)$ in the original mechanism.

Generally speaking, the strategy costs can be thought of as financial costs or efforts paid by agents when carrying out their strategies. According to MWG's textbook (Page 883, Line 7 [1]), agents' strategies are either possible actions or plans of actions. No matter which format the agents' strategies might be, if the strategy costs occurred in the original mechanism cannot be ignored, then only when such assumption holds will the mediator's announcement be credible to the agents. Otherwise none of agents is willing to attend the direct mechanism, which means the direct mechanism cannot start up.

However, the so-called "mediator" is a virtual role which does not exist at all. It is *unreasonable* to assume that the nonexistent"mediator" pay the strategy costs for agents. Hence, the above explanation of the revelation principle using the virtual "mediator" is *wrong* when agents' strategies are costly. According to Definition 23.B.5 (See Appendix), the only *legal* action for each agent $i$ in the direct mechanism is just to report a type $\theta_i$ (which means that $s_i^*(\theta_i)$ is illegal for agent $i$), and the outcome function is just the social choice function. From the perspective of agents, the above-mentioned result is formalized as the following strategy cost property:

**Strategy cost property**: *In the direct mechanism, each agent only reports a type and will not pay the strategy cost which would be paid by himself when playing strategies in the original indirect mechanism.*

The strategy cost property can be understood by proof of contradiction. Reconsider the example in Section 2. Suppose there are two different kinds of certifications from which the firm can choose one as a hire criterion, and the

education levels corresponding to the two certifications are different: $e'_H$ and $e''_H$. Hence, after the firm announces a certification, by Eq (3) and Eq (7) each agent in the indirect mechanism will know which kind of strategy cost he should pay in Bayesian equilibrium: $e'_H/\theta_H$ or $e''_H/\theta_H$. Now assume that each agent will still pay the strategy cost in the direct mechanism. Then after each agent reports a type to the firm, the firm performs the outcome function which is just the social choice function $f$. Since there is no certification used in the direct mechanism, it is impossible for each agent $i$ to know which kind of strategy cost he should pay: $e'_H/\theta_H$ or $e''_H/\theta_H$. This is the contradiction. Consequently, the strategy cost property holds.

One possible question to the strategy cost property is as follows:
*Q1:* The designer may define the direct mechanism more generally. In particular, The designer defines a new mechanism in which each agent reports his type, then the mechanism suggest to them which action to take, and the final outcome of the mechanism depends on both the report and the action (*i.e.*, education level).
*A1:* This new mechanism is irrelevant to the direct mechanism. By Definition 23.B.5 (See Appendix), the final outcome of the direct mechanism only depends on agents' reports, and the designer must perform the outcome function after receiving agents' reports. It is *illegal* to assume that the designer can send action advices to agents in the direct mechanism.

Besides the above question, another possible objection to the strategy cost property is as follows: "Let us consider the equilibrium in the indirect mechanism. Given the equilibrium, there is a mapping from vectors of agents' types into outcomes. Now let us take that mapping to be a revelation game. It will be the case that no type of any agent can make an announcement that differs from his true type and do better".

It can be seen that this objection is equivalent to the proof of revelation principle (see Appendix Proposition 23.D.1). Suppose the strategy costs cannot be neglected in the indirect mechanism, let us make a detailed investigation on the proof of Proposition 23.D.1. Given that an indirect mechanism $\Gamma$ implements $f$ costly in Bayesian Nash equilibrium, consider the equilibrium $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ in Eq (23.D.2), there is a mapping $g(s^*(\cdot)) : \Theta_1 \times \cdots \times \Theta_I \to X$ from a vector of agents' types $\theta = (\theta_1, \cdots, \theta_I)$ into an outcome $g(s^*(\theta))$, which is equal to the desired outcome $f(\theta)$ for all $\theta \in \Theta_1 \times \cdots \times \Theta_I$. Note that in Eq (23.D.2) and Eq (23.D.3), the indirect mechanism $\Gamma$ works, and the utility function $u_i$ of agent $i$ ($i = 1, \cdots, I$) already reflects the fact that each agent $i$ pays the strategy cost related to $s_i^*(\theta_i)$ by himself.

Now let us take the mapping $g(s^*(\cdot))$ to be a direct revelation game, in which the strategy set of agent $i$ is his type set, $S_i = \Theta_i$, and the designer carries out the outcome function $f(\cdot)$. In this revelation game, each agent $i$ only reports a

type and does not pay the strategy cost except for some possible misreporting cost, thus *the utility function of each agent $i$ in the direct mechanism should be changed from original $u_i$ to another function $u_i'$, in which the item related to strategy cost disappears.* [2]

As a result, given that an indirect mechanism implements $f$ costly in Bayesian Nash equilibrium, in order to judge whether $f$ is truthfully implementable in Bayesian Nash equilibrium or not, we should use the new utility function $u_i'$ instead of $u_i$ for each agent $i$. To be more precisely, the criterion to judge whether $f$ is truthfully implementable should be updated from Eq (23.D.1) (See Appendix) to judge whether for all $i = 1, \cdots, I$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i'(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i'(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i], \tag{9}$$

for all $\hat{\theta}_i \in \Theta_i$, in which $u_i'$ is the utility function of agent $i$ in the direct mechanism, and is not equal to $u_i$ in the original indirect mechanism.

Therefore, the last sentence of the proof of Proposition 23.D.1 (See Appendix) is *wrong* since Eq (23.D.4) is no longer the condition for $f$ to be truthfully implementable in Bayesian Nash equilibrium when strategies in the indirect mechanism are costly. Furthermore, with the new utility function $u_i'$, some agent $i$ may find it beneficial for him to differ from his true type $\theta_i$ to another false type $\hat{\theta}_i$. [3] Put differently, in the direct mechanism there may exists some agent $i \in \{1, \cdots, I\}$, $\theta_i, \hat{\theta}_i \in \Theta_i$ such that

$$E_{\theta_{-i}}[u_i'(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] < E_{\theta_{-i}}[u_i'(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i],$$

which contradicts Eq (9).

To sum up, the strategy cost property is the cornerstone for the direct revelation mechanism to start up. However, as we pointed out, *it is the strategy cost property itself that may change agents' utility functions in the direct mechanism*. Consequently, a costly Bayesian implementable social choice function may not be truthfully implementable, which contradicts the revelation principle. Note that in this section, only the strategy costs are considered and the misreporting costs has nothing to do with the conclusion. Put differently, the revelation principle may not hold even if the misreporting costs are zero and only the agents' strategies are costly. An example will be shown in Section 5.

---

[2] In Section 5, the utility function of agent $i$ in the direct mechanism is changed from Eq (3) to Eq (10) and Eq (11), in which the item related to the strategy cost "$b_i/\theta_i$" disappears.

[3] In Section 5, each worker $i = 1, 2$ finds it beneficial to misreport $\hat{\theta}_i = \theta_H$ in the direct mechanism under the condition of $c' \in (0, 0.5w)$, no matter what his true type is.

## 5 $f$ is not truthfully implementable in Bayesian Nash equilibrium

**Proposition 2:** If the misreporting cost $c' \in (0, 0.5w)$, the social choice function $f(\theta)$ given in Eq (1) is not truthfully implementable in Bayesian Nash equilibrium.

**Proof:** Consider the direct revelation mechanism $\Gamma_{direct} = (\Theta_1, \Theta_2, f(\theta))$, in which $\Theta_1 = \Theta_2 = \{\theta_L, \theta_H\}$, $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$. The timing steps of $\Gamma_{direct}$ are as follows:

1) Each worker $i$ $(i = 1, 2)$ with true type $\theta_i$ reports a type $\hat{\theta}_i \in \Theta_i$ to the firm. Here $\hat{\theta}_i$ may not be equal to $\theta_i$.

2) The firm performs the outcome function $f(\hat{\theta}_1, \hat{\theta}_2)$, and hires the winner.

According to the strategy cost property, in the direct mechanism, each worker $i$ only reports a type and does not pay the strategy cost. The only possible cost needed to pay is the misreporting cost $c'$ for a low-productivity worker to report the high productivity type $\theta_H$. For worker $i$ $(i = 1, 2)$, if his true type is $\theta_i = \theta_L$, his utility function will be as follows:

$$u_i'(\hat{\theta}_i, \hat{\theta}_j; \theta_i = \theta_L) = \begin{cases} w - c', & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_L) \\ 0.5w - c', & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H) \\ 0.5w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_L) \\ 0, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_H) \end{cases}, \ i \neq j. \tag{10}$$

If worker $i$'s true type is $\theta_i = \theta_H$, his utility function will be as follows:

$$u_i'(\hat{\theta}_i, \hat{\theta}_j; \theta_i = \theta_H) = \begin{cases} w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_L) \\ 0.5w, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H), or (\theta_L, \theta_L) \\ 0, & \text{if } (\hat{\theta}_i, \hat{\theta}_j) = (\theta_L, \theta_H) \end{cases}, \ i \neq j. \tag{11}$$

Note that the strategy cost item "$b_i/\theta_i$" occurred in Eq (3) disappears in Eq (10) and Eq (11). Following Eq (10) and Eq (11), we will discuss the utility matrix of worker $i$ and $j$ in four cases. The first and second entry in the parenthesis denote the utility of worker $i$ and $j$ respectively.

1) Suppose the true types of worker $i$ and $j$ are $\theta_i = \theta_H$, $\theta_j = \theta_H$.

| $\hat{\theta}_i$ \ $\hat{\theta}_j$ | $\theta_L$ | $\theta_H$ |
|---|---|---|
| $\theta_L$ | $(0.5w, 0.5w)$ | $(0, w)$ |
| $\theta_H$ | $(w, 0)$ | $(0.5w, 0.5w)$ |

It can be seen that: the dominant strategy for worker $i$ and $j$ is to truthfully report, *i.e.*, $\hat{\theta}_i = \theta_H$, $\hat{\theta}_j = \theta_H$. Thus, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

2) Suppose the true types of worker $i$ and $j$ are $\theta_i = \theta_L$, $\theta_j = \theta_H$.

| $\hat{\theta}_i$ \ $\hat{\theta}_j$ | $\theta_L$ | $\theta_H$ |
|---|---|---|
| $\theta_L$ | $(0.5w, 0.5w)$ | $(0, w)$ |
| $\theta_H$ | $(w - c', 0)$ | $(0.5w - c', 0.5w)$ |

It can be seen that: the dominant strategy for worker $j$ is still to truthfully report $\hat{\theta}_j = \theta_H$; and if the misreporting cost $c' < 0.5w$, the dominant strategy for worker $i$ is to misreport $\hat{\theta}_i = \theta_H$, otherwise agent $i$ would truthfully report. Thus, under the condition of $c' \in (0, 0.5w)$, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

3) Suppose the true types of worker $i$ and $j$ are $\theta_i = \theta_H$, $\theta_j = \theta_L$.

| $\hat{\theta}_i$ \ $\hat{\theta}_j$ | $\theta_L$ | $\theta_H$ |
|---|---|---|
| $\theta_L$ | $(0.5w, 0.5w)$ | $(0, w - c')$ |
| $\theta_H$ | $(w, 0)$ | $(0.5w, 0.5w - c')$ |

It can be seen that: the dominant strategy for worker $i$ is still to truthfully report $\hat{\theta}_i = \theta_H$; and if the misreporting cost $c' < 0.5w$, the dominant strategy for worker $j$ is to misreport $\hat{\theta}_j = \theta_H$, otherwise agent $j$ would truthfully report. Thus, under the condition of $c' \in (0, 0.5w)$, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

4) Suppose the true types of worker $i$ and $j$ are $\theta_i = \theta_L$, $\theta_j = \theta_L$.

| $\hat{\theta}_i$ \ $\hat{\theta}_j$ | $\theta_L$ | $\theta_H$ |
|---|---|---|
| $\theta_L$ | $(0.5w, 0.5w)$ | $(0, w - c')$ |
| $\theta_H$ | $(w - c', 0)$ | $(0.5w - c', 0.5w - c')$ |

It can be seen that: if the misreporting cost $c' < 0.5w$, the dominant strategy for both worker $i$ and worker $j$ is to misreport, *i.e.*, $\hat{\theta}_i = \theta_H$, $\hat{\theta}_j = \theta_H$, otherwise both agents would truthfully report. Thus, under the condition of $c' \in (0, 0.5w)$, the unique Nash equilibrium is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$.

To sum up, under the condition of $c' \in (0, 0.5w)$, the unique Nash equilibrium of the game induced by the direct mechanism is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$, and the unique outcome of $\Gamma_{direct}$ is that each worker has the same probability 0.5 to get the job offer. Consequently, $\hat{\theta}_i^* = \theta_i$ (for all $\theta_i \in \Theta_i$, $i = 1, 2$) is not a Bayesian Nash equilibrium of the direct revelation mechanism under the condition of $c' \in (0, 0.5w)$, and hence the social choice function $f(\theta)$ is not truthfully implementable in Bayesian Nash equilibrium. □

## 6 Conclusions

In this paper, we discuss the justification of revelation principle through a simple labor model in which agents pay strategy costs during the process of an indirect mechanism. The main characteristics of the labor model are as follows: 1) Agents' strategies are costly in the indirect mechanism, *i.e.*, worker with type $\theta_H$ (or $\theta_L$) will pay the strategy cost $e_H/\theta_H$ (or $e_H/\theta_L$) when obtaining education level $e_H$; 2) The productivity type of worker is private information and not observable to the firm; 3) Misreporting a higher type is also costly, *i.e.*, a low-productivity worker can pretend to be a high-productivity worker with the misreporting cost $c'$.

The major difference between this paper and traditional literature is just the strategy cost property proposed in Section 4. By the strategy cost property, when strategies in the indirect mechanism are costly, the utility function of agents will be changed in the direct mechanism. Hence, the criterion to judge whether $f$ is truthfully implementable in Bayesian Nash equilibrium will also be changed.

Section 3 and Section 5 give detailed analysis about the labor model:
1) In the indirect mechanism $\Gamma$, the utility function of each worker $i = 1, 2$ is given by Eq (3), in which the strategy cost $b_i/\theta_i$ is the key item that makes the separating strategy profile $(b_1^*(\theta_1), b_2^*(\theta_2))$ be a Bayesian Nash equilibrium if the wage $w \in (2e_H/\theta_H, 2e_H/\theta_L)$. Thus, the social choice function $f$ can be implemented in Bayesian Nash equilibrium.
2) Following the strategy cost property, in the direct mechanism, the utility function of each worker $i$ is changed from Eq (3) to Eq (10) and Eq (11). Under the condition of $c' \in (0, 0.5w)$, the unique Nash equilibrium of the game induced by the direct mechanism is $(\hat{\theta}_i, \hat{\theta}_j) = (\theta_H, \theta_H)$, and $\hat{\theta}_i^* = \theta_i$ (for all $\theta_i \in \Theta_i$, $i = 1, 2$) is not a Bayesian Nash equilibrium of the direct mechanism. Thus, the social choice function $f$ is not truthfully implementable in Bayesian Nash equilibrium.

In summary, the revelation principle may not hold when agents' strategies are costly in the indirect mechanism.

## Appendix: Definitions in Section 23.B and 23.D [1]

Consider a setting with $I$ agents, indexed by $i = 1, \cdots, I$. Each agent $i$ privately observes his type $\theta_i$ that determines his preferences. The set of possible types of agent $i$ is denoted as $\Theta_i$. The agent $i$'s utility function over the outcomes in set $X$ given his type $\theta_i$ is $u_i(x, \theta_i)$, where $x \in X$.

**Definition 23.B.1**: A *social choice function* is a function $f : \Theta_1 \times \cdots \times \Theta_I \to X$ that, for each possible profile of the agents' types $(\theta_1, \cdots, \theta_I)$, assigns a collective choice $f(\theta_1, \cdots, \theta_I) \in X$.

**Definition 23.B.3**: A *mechanism* $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ is a collection of $I$ strategy sets $S_1, \cdots, S_I$ and an outcome function $g : S_1 \times \cdots \times S_I \to X$.

**Definition 23.B.5**: A *direct revelation mechanism* is a mechanism in which $S_i = \Theta_i$ for all $i$ and $g(\theta) = f(\theta)$ for all $\theta \in \Theta_1 \times \cdots \times \Theta_I$.

**Definition 23.D.1**: The strategy profile $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ is a *Bayesian Nash equilibrium* of mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ if, for all $i$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i]$$

for all $\hat{s}_i \in S_i$.

**Definition 23.D.2**: The mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ *implements the social choice function* $f(\cdot)$ *in Bayesian Nash equilibrium* if there is a Bayesian Nash equilibrium of $\Gamma$, $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$, such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

**Definition 23.D.3**: The social choice function $f(\cdot)$ is *truthfully implementable in Bayesian Nash equilibrium* if $s_i^*(\theta_i) = \theta_i$ (for all $\theta_i \in \Theta_i$) is a Bayesian Nash equilibrium of the direct revelation mechanism $\Gamma = (\Theta_1, \cdots, \Theta_I, f(\cdot))$. That is, if for all $i = 1, \cdots, I$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i], \qquad (23.D.1)$$

for all $\hat{\theta}_i \in \Theta_i$.

**Proposition 23.D.1**: (*The Revelation Principle for Bayesian Nash Equilibrium*) Suppose that there exists a mechanism $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ that implements the social choice function $f(\cdot)$ in Bayesian Nash equilibrium. Then $f(\cdot)$ is truthfully implementable in Bayesian Nash equilibrium.
**Proof:** If $\Gamma = (S_1, \cdots, S_I, g(\cdot))$ implements $f(\cdot)$ in Bayesian Nash equilibrium, then there exists a profile of strategies $s^*(\cdot) = (s_1^*(\cdot), \cdots, s_I^*(\cdot))$ such that $g(s^*(\theta)) = f(\theta)$ for all $\theta$, and for all $i$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i], \quad (23.D.2)$$

for all $\hat{s}_i \in S_i$. Condition (23.D.2) implies, in particular, that for all $i$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i], \quad (23.D.3)$$

for all $\hat{\theta}_i \in \Theta_i$. Since $g(s^*(\theta)) = f(\theta)$ for all $\theta$, (23.D.3) means that, for all $i$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i], \quad (23.D.4)$$

for all $\hat{\theta}_i \in \Theta_i$. But, this is precisely condition (23.D.1), the condition for $f(\cdot)$ to be truthfully implementable in Bayesian Nash equilibrium. □


## Acknowledgments

The author is grateful to Fang Chen, Hanyue, Hanxing and Hanchen for their great support.


## References


[1] A. Mas-Colell, M.D. Whinston and J.R. Green, Microeconomic Theory, Oxford University Press, 1995.

[2] A. Kephart and V. Conitzer, The revelation principle for mechanism design with reporting costs, In Proceedings of the ACM Conference on Electronic Commerce (EC), Maastricht, The Netherlands, 2016.

[3] J. Green and J.J. Laffont, Partially verifiable information and mechanism design. Review of Economic Studies, vol.53, 447-456, 1986.

[4] L. Yu, Mechanism design with partial verification and revelation principle. Autonomous Agents and Multi-Agent Systems, vol.22, 217-223, 2011.

[5] V. Auletta, *et al*, Alternatives to truthfulness are hard to recognize. Autonomous Agents and Multi-Agent Systems, vol.22, 200-216, 2011.

[6] M. Spence, Job Market Signaling. Quarterly Journal of Economics, vol.87, 355-374, 1973.