G-COE GLOPE II Working Paper Series

Theoretical and experimental investigations of the performance of keyword auction mechanisms

Emiko Fukuda, Yoshio Kamijo, Ai Takeuchi, Michiharu Masui,

and Yukihiko Funaki

Working Paper No. 33

Theoretical and experimental investigations of the performance of keyword auction mechanisms*

Emiko Fukuda[a], Yoshio Kamijo[b], Ai Takeuchi[c]†, Michiharu Masui[d], and Yukihiko Funaki[e]

[a] Department of Computer Science, National Defense Academy of Japan. 1-10-20 Hashirimizu, Yokosuka, Kanagawa 239-8686, JAPAN

[b] Waseda Institute for Advanced Study (WIAS). 1-6-1, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, JAPAN

[c] Faculty of Political Science and Economics, Waseda University. 1-6-1, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, JAPAN

[d] Graduate School of Decision Science and Technology, Tokyo Institute of Technology. 2-12-1 Oh-okayama, Meguro-ku, Tokyo 152-8552, JAPAN

[e] School of Political Science and Economics, Waseda University. 1-6-1, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, JAPAN

December 2010

Abstract

Two keyword auction mechanisms, the Generalized Second-Price auction (GSP) and the Vickrey-Clarke-Groves mechanism (VCG), were compared theoretically and experimentally. The former is widely used in practice, while the latter is not used in practice, but has a dominant strategy equilibrium where all the participants bid their true values. In the theoretical investigation, by applying the "locally envy-free Nash equilibrium" to the VCG, we found that the allocations are efficient and that upper and lower bounds of the auctioneer's revenue among all the locally envy-free Nash equilibria coincide in the two mechanisms. This theoretical result was supported in a laboratory experiment, in which the revenues and efficiencies were similar in both mechanisms. In addition, Nash equilibrium and locally envy-free Nash equilibrium bid profiles were more frequently observed in the VCG than in the GSP.

Keywords: Keyword Auction, Generalized Second-Price Auction, Vickrey-Clarke-Groves Mechanism, Laboratory Experiment
JEL Codes: C72, C91, D44

*We are grateful to John Wooders for helpful comments and suggestions, which significantly improved the paper. We thank Tatsuhiro Shichijo, Takehiko Yamato, and seminar participants at the 5th Spain, Italy, Netherlands Meeting on Game Theory (SING 5) and the 2009 Far East and South Asia Meeting of the Econometric Society (FESAMES 2009) for helpful comments and discussion. This experiment was supported by the Grant-in-Aid for the Global COE "Political Economy of Institutional Construction" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

†Corresponding Author: Tel:+81 3 3442 8827, Fax:+81 3 3442 8827. E-mail Addresses: ai-tak@moegi.waseda.jp (A. Takeuchi); emiko@nda.ac.jp (E. Fukuda); yoshio.kamijo@gmail.com (Y. Kamijo); masui.m.aa@m.titech.ac.jp (M. Masui); funaki@waseda.jp (Y. Funaki).

1 Introduction

In keyword auctions, positions for Internet advertisements are sold.[1] These auctions are conducted by search engines such as Google, Yahoo!, and MSN and have contributed to the large commercial success of the Internet industry. The advertisement revenue from these keyword auctions has increased considerably over the last few years. For example, Google's revenue from these auctions increased more than 2.5-fold from $6 billion in 2005 to $16.4 billion in 2007. Since 2003, the revenue generated from keyword auctions has accounted for approximately 47% of the total Internet advertisement revenue in the USA, which increased from $7.2 billion in 2003 to $22.7 billion in 2009.[2] Similarly in Japan, the revenue from these auctions increased from 59 billion Yen (about $500 million) in 2005 to 171 billion Yen (about $1.8 billion) in 2009.[3]

The process behind keyword auctions is as follows. Advertisements are displayed along with the search results of a query submitted to a search engine, and clicking on the advertisement leads the users to the advertiser's webpage. The users obtain information on the goods or services offered on the webpage and can choose to purchase them, generating revenue for the advertiser. Thus, the total revenue achieved by an advertiser is determined by the number of clicks on the advertisement. The number of clicks depends on the position of the advertisement, termed as ad-spot. Advertisements in higher ad-spots tend to be clicked more frequently. Therefore, aside from the payments to the search engine, the total profit for advertisers will be higher if they obtain higher ad-spots. Ad-spot allocations and payments to the search engines are determined by keyword auctions. Each keyword is sold through a separate auction, so advertisers can place bids for a keyword or several keywords that when searched for, would trigger the display of their advertisements. Since there are numerous possible keywords, hundreds of millions of keyword auctions are conducted every day. The properties of auction mechanisms for keyword auctions have begun to attract the attention of economists, computer scientists, and advertisers. In this study, we compare two auction mechanisms for keyword auctions theoretically and experimentally.

Keyword auctions are conducted mostly through Generalized Second-Price auction (GSP). Their use has contributed to the outstanding success of the search engines mentioned above, and they are attracting the most attention among keyword auction mechanisms. The GSP mechanism is simple. First, the ad-spots are allocated in the descending order of the advertisers' bids, i.e., the top position is allocated to the bidder with the highest bid, the second position is allocated to the bidder with the second highest bid, and so on.[4] Advertisers pay an amount equal to the next highest bid every time a search engine user clicks the bidder's advertisement. When there is only one ad-spot, the GSP is equivalent to the second-price auction; therefore, the GSP is a generalization of the second-price auction to the environment of keyword auctions where there are multiple ad-spots. The GSP has evolved from

[1]Keyword auctions are also referred to as the sponsored search auction, position auction, and ad auction. Also in the literature, the advertisements displayed are sometimes called sponsored links or ad-words.

[2]These data are from annual reports by the Interactive Advertising Bureau (http://www.iab.net/insights_research/947883/adrevenuereport, last searched on October 2010.) To be more precise, these percentages correspond to a category called "Search," which consists of revenues not only from keyword auctions, but also from other forms of Internet advertisements that allow for the specification of keywords to have advertisements displayed. The percentage of revenues solely from keyword auctions could not be computed from these reports.

[3]These data are from annual reports by Dentsu, a leading advertisement company in Japan (http://www.dentsu.co.jp/marketing/koukokudata.html, last searched on October 2010; the pages are in Japanese.) A similar argument as that in Footnote 2 applies.

[4]Presently, both Yahoo! and Google adopt a modification of the GSP, in which ad-spots are allocated in the decreasing order of the ranking scores that are determined for each advertiser i as the product of i's bid and the estimated quality score of i (see Aggarwal et al., 2006A, 2006B for detail.) The GSP we investigated in this study, like the one researched by many other studies, is equivalent to the case where the quality scores of all advertisers are equal to 1. This excludes one variable and simplifies the experiment setting, which allows for a more controlled comparison of the mechanisms.

the generalized first-price auction, where advertisers pay the bid that they submit, and has remained the dominant mechanism for keyword auctions ever since (see Edelman et al. (2007) for a detailed explanation of the transition of the auction mechanism actually used by search engines).

An alternative auction mechanism for keyword auctions is the Vickrey-Clarke-Groves mechanism (VCG). The VCG has a nice theoretical property that bidding one's own true value is the dominant strategy; however, it is not used in practice. When there is only one ad-spot, the VCG is equivalent to the second-price auction; thus, it is equivalent to the GSP. However, while truth-telling is always the dominant strategy in the VCG, the GSP does not have a dominant strategy equilibrium, and submitting one's true value may not even be a Nash equilibrium when multiple objects are for sale. Edelman and Ostrovsky (2007) reported that bids in the GSP are unstable because of bidders' strategic behaviors. In contrast, the bids and revenues of the VCG might be more stable. Therefore, the VCG is an alternative auction system that may reduce the problem of bidders' strategic behaviors and improve the auctioneer's revenue.

Theoretical results from the existing literature indicate several similarities and difference between the two mechanisms; efficiency would be achieved by both mechanisms, while revenues for the auctioneer would be higher with the GSP (Edelman et al., 2007 and Varian, 2007). Each study showed that the revenues for the auctioneer of the GSP in newly defined locally envy-free Nash equilibria (LEFNE)–Nash equilibria where no bidder is better off by exchanging his or her position with that of the bidder immediately above–are greater than or equal to the revenue in the dominant strategy equilibrium of the VCG. Therefore, under the assumption that the LEFNE and the dominant strategy equilibrium are achieved in the GSP and the VCG, respectively, the GSP is equally or more profitable for the auctioneer than the VCG. They also showed that at any LEFNE of the GSP, the allocations of the ad-spots are efficient. Because the allocation in the dominant strategy equilibrium of the VCG is efficient, the two auction systems achieve the same performance in this respect.

In this study, we first re-examine the theoretical properties of the two mechanisms explored in the existing literature. So far, the dominant strategy equilibrium has been used to analyze the VCG. However, a number of experiments on the VCG report that the dominant strategy is rarely chosen by subjects. For example, Kagel et al. (1987), Kagel and Levin (1993), and Harstad (2000) report experiments on the second-price auction where subjects did not bid their true values. Therefore the dominant strategy equilibrium is not necessarily a valid concept for predicting the performance of the mechanisms. In addition, because the dominant strategy equilibrium is unique and there are multiple LEFNE, it is difficult to compare performance based on these equilibrium concepts where one has a range and the other is unique. Therefore, it is necessary to re-examine the theoretical performance of the two mechanisms by using the same equilibrium concept. We applied the LEFNE to the VCG and obtained the following results. First, in any LEFNE of the VCG, the ad-spot allocations are efficient. Second, in the VCG, the dominant strategy equilibrium is a LEFNE and generates minimal revenue among all the LEFNE. Third, the maximum revenue among all LEFNE of the VCG is equal to that of the GSP. The second and third results, together with the aforementioned results by Edelman et al. and Varian, imply that not only the lower bound of the revenues coincides in the GSP and the VCG, but also the upper bound of the revenues coincides in the two mechanisms. Therefore, by theoretically comparing the two auction mechanisms, we find that the VCG and the GSP are similar in both efficiency and profitability and that we cannot judge either mechanism's superiority in these respects.

Although these results suggest no difference in the performance of the two mechanisms, there is no guarantee that if actually implemented, the results of these auction mechanisms will be the same, because the assumptions behind the theory may not model an actual agent's behavior. Therefore, we conducted a laboratory experiment to compare the performance of the two mechanisms. Many simulations and empirical analyses have been conducted for keyword auctions, especially on the GSP (e.g., Cary et al., 2007; Edelman and Ostrovsky, 2007; Varian, 2007, 2009). Analyses based on experiments have two advantages. First, it is possible to analyze a mechanism that does not yet exist in practice. Second, we can specify the values of the advertisers. The second advantage allows us to directly analyze the frequency of equilibrium and to calculate surplus. Because of these advantages, many experiments have been conducted in various environments and on several mechanisms (see Kagel, 1995; Kagel and Levin, 2008 for surveys on experiments on auctions), but no experiments have yet examined keyword auctions. This is the first experimental investigation to compare the GSP and the VCG.[5]

The keyword auctions are generally continuous and dynamic in practice, so the advertisers can change their bids any time. In the theoretical part of the analysis, the game is modeled as a one-shot game. As discussed in Edelman and Schwarz (2010), a well-modeled static game captures important points of the dynamic game while keeping the analysis tractable, and there is a long history in modeling the dynamic game of incomplete information as a one-shot game of complete information (e.g., Cournot model and Bertrand model). The theoretical analysis followed these approaches. In the experiment, however, we were more interested in comparing the two mechanisms in a dynamic game of incomplete information, which is closer to a keyword auction in practice. The game played in the experiment is thus the dynamic game of incomplete information. With this design we will not be able to reject the theory even if the experimental results reject them. However, if the experimental results are in line with the theoretical results based on the model of the one-shot game, they will provide support for the theory.

Experimental observations were as follows. First, although the relative frequencies of the equilibrium bids were low in both mechanisms, the number of Nash equilibria and the LEFNE were higher in the VCG than in the GSP. Next, we observed similar levels of efficiencies and the auctioneer's revenues in the two mechanisms. Furthermore, in both the VCG and the GSP, the revenues were close to the lower bound. In sum, our experimental results on revenue and efficiency were in line with the theoretical prediction based on the LEFNE.

This paper is organized as follows. In Section 2, we review the definitions and properties of the GSP and the VCG, and provide a theoretical analysis. The experimental procedures are outlined in Section 3, and the results of the experiment are reported in Section 4. We conclude by summarizing the results in Section 5.

2 Model

2.1 The GSP and the VCG

A keyword auction has the following components. There are N advertisers (bidders) participating in a keyword auction. Each advertiser i has a value or expected revenue v_i for a click of the ad. We assume that $v_1 > v_2 > ... > v_N$. There are K ad-spots with click-through rates (CTRs) $\alpha_1 > \alpha_2 > ... > \alpha_K > 0$, where α_k is the estimated probability of being clicked or the estimated number of clicks per given period, for an advertiser in the k-th ad-spot. We also set $\alpha_k = 0$ for all $k > K$ and assume $N = K$. Each advertiser submits a bid to the auction. The bid submitted by i is denoted by b_i. We denote the bid profile of N advertisers by $b = (b_1, \dots, b_N)$. An auction mechanism determines how ad-spots are assigned to advertisers and the cost of advertisers for each click, depending on their bid profile.

Literature in this field pays a great deal of attention to two types of mechanisms—the GSP and the VCG. In the GSP, ad-spots are allocated to advertisers in the descending order of their bids $b_1, b_2, ..., b_N$. Let $d(k)$ denote the name of the bidder who submits the k-th highest bid among $b_1, b_2, ..., b_N$. Thus, bidder $d(k)$ acquires ad-spot k. The advertiser obtaining ad-spot k pays the bid of the advertiser obtaining

[5] Ostrovsky and Schwarz (2009) conducted a large field experiment, manipulating the reserve price in Yahoo!'s keyword auctions.

one lower ad-spot for each click. Thus, for each ad-spot k, advertiser $d(k)$ pays $b_{d(k+1)}$ for a click of its ad and the payment $p_k^G(b)$ of the bidder in ad-spot k is $\alpha_k b_{d(k+1)}$. Note that in our setting, $d(K+1)$ is not well defined, and thus, we set $b_{d(K+1)} = 0$ for the completeness of the payments of advertisers. The payoff of the advertiser obtaining ad-spot k is $\alpha_k v_{d(k)} - p_k^G(b) = \alpha_k(v_{d(k)} - b_{d(k+1)})$.

When there is only one ad-spot, the GSP becomes the usual second-price auction for a single object because the winner of the ad-spot pays the second largest bid for a click. The second-price auction is known as a truthful mechanism; bidding the true value of a sale item is the dominant strategy for each bidder. However, when there are multiple ad-spots, the GSP is not truthful and has multiple Nash equilibria. Thus, advertisers participating in the GSP face some difficulties when determining their bids. In contrast, the VCG is known to be a truthful mechanism.

In the VCG for keyword auctions, the ad-spots are assigned to the advertisers in the same manner as in the GSP. For each ad-spot k, advertiser $i = d(k)$ pays the negative externality that i imposes on other advertisers. The payment of advertiser i who acquires ad-spot k is

$$p_k^V(b) = \left[\sum_{h=1}^{k-1} \alpha_h b_{d(h)} + \sum_{h=k+1}^{K} \alpha_{h-1} b_{d(h)}\right] - \left[\sum_{h=1}^{k-1} \alpha_h b_{d(h)} + \sum_{h=k+1}^{K} \alpha_h b_{d(h)}\right], \tag{1}$$

where, assuming that each advertiser submits their true value for a click, the expression in the first square bracket is the sum of the revenue of advertisers other than i when i leaves the auction, and the expression in the second square bracket is the sum of their revenues when i participates in the auction.

By simple calculation, Eq. (1) is reduced to

$$p_k^V(b) = \sum_{h=k+1}^{K} (\alpha_{h-1} - \alpha_h) b_{d(h)} = \sum_{h=k}^{K-1} (\alpha_h - \alpha_{h+1}) b_{d(h+1)}.$$

From this, we obtain a convenient recursive formula for the payment of advertisers: $p_K^V(b) = 0$, and for each $k = 1, 2, \ldots, K-1$, $p_k^V(b) = (\alpha_k - \alpha_{k+1}) b_{d(k+1)} + p_{k+1}^V(b)$.

2.2 A locally envy-free Nash equilibrium (LEFNE)

In the VCG, truth-telling is the dominant strategy for each advertiser. In contrast, the GSP is not a truth-telling mechanism and has multiple Nash equilibria. Therefore, most of the papers on keyword auctions focus on a refined concept, called a LEFNE.

The bid profile b is a LEFNE if the following two conditions hold: for each $k = 2, ..., K$,

$$\alpha_k v_{d(k)} - p_k^G(b) \geq \alpha_{k-1} v_{d(k)} - p_{k-1}^G(b), \tag{2}$$

and, for each $k = 1, ..., K-1$,

$$\alpha_k v_{d(k)} - p_k^G(b) \geq \alpha_{k+1} v_{d(k)} - p_{k+1}^G(b). \tag{3}$$

These conditions require that in their current state, no bidder is better off by exchanging their position with that of the bidder immediately above (Inequality (2)) and immediately below (Inequality (3)). For i with $i = d(k)$, if Inequalities (2) and (3) hold for this k, we say that b_i is LEF at the bid profile b.

If we replace p_k^Gs in Inequalities (2) and (3) with p_k^Vs, we obtain the definition of the LEFNE for the VCG.

In the literature, a LEFNE is introduced with slightly different definitions. Edelman et al. (2007) called a bid profile a locally envy-free equilibrium if it is a Nash equilibrium of a one-shot GSP auction and satisfies Inequality (2). Varian (2007) introduced an equilibrium concept that requires that in the current bid profile, no bidder should be better off by exchanging their position with that of any other bidder.[6] Interestingly, it is now known that a LEFNE defined by (2) and (3) is equivalent to the equilibrium concepts introduced in Edelman et al. (2007) and Varian (2007).

2.3 Theoretical results

The following proposition lists some of the theoretical properties of the GSP mechanism that were explored by Edelman et al. (2007) and Varian (2007).

Proposition 1 (Edelman et al., 2007 and Varian, 2007)**.** *The following statements about the GSP auction hold true.*

(i) The auctioneer's revenue (the sum of the payments of all advertisers) in the GSP in any LEFNE is greater than or equal to the auctioneer's revenue in the VCG in the dominant strategy equilibrium.

(ii) In any LEFNE of the GSP, the allocation of the ad-spots is efficient.

(iii) The maximum revenue among LEFNE is achieved in the following LEFNE bid profile $\bar{b}^G$:

$$\bar{b}^G_k = \frac{1}{\alpha_k}\sum_{h=k}^{K} v_{h-1}(\alpha_{h-1} - \alpha_h) \text{ for each } k = 2, 3, ..., N \text{ and any } \bar{b}^G_1 \text{ with } \bar{b}^G_1 > \bar{b}^G_2.$$

(iv) The minimum revenue among LEFNE is achieved in the following LEFNE bid profile $\underline{b}^G$:

$$\underline{b}^G_k = \frac{1}{\alpha_k}\sum_{h=k}^{K} v_h(\alpha_{h-1} - \alpha_h) \text{ for each } k = 2, 3, ..., N \text{ and any } \underline{b}^G_1 \text{ with } \underline{b}^G_1 > \underline{b}^G_2.$$

The first statement of this proposition is one of the theoretical answers to the question of why the GSP is used in practice but the VCG is not. However, there is neither theoretical support nor empirical evidence that the LEFNE is achieved in the actual auction.[7] This means that the revenue gained by the auctioneer in the GSP may not be greater than that gained in the VCG if we test their performances in a laboratory experiment. The second statement of this proposition means that a LEFNE is sufficient to guarantee the efficient allocation of the ad-spots. In the context of a keyword auction, a bid profile is efficient if it is in assortative order ($b_1 > b_2 > ... > b_N$). The third and fourth statements show that the bid profiles that achieve the maximum and minimum revenues of the GSP are easily calculated. From these two, the maximum revenue R_M and the minimum revenue R_m of the GSP are obtained as follows:

$$R_M = \sum_{k=1}^{K-1} k v_k(\alpha_k - \alpha_{k+1}), \tag{4}$$

and

$$R_m = \sum_{k=1}^{K-1} k v_{k+1}(\alpha_k - \alpha_{k+1}). \tag{5}$$

[6] Varian (2007) called this concept a symmetric Nash equilibrium.

[7] There are several evidences of the validity of the LEFNE in the computer science literatures. Cary et al. (2007) examine the greedy bidding strategy called balanced bidding in the repeatedly played GSP auction, and show that if in each period, only one bidder changes their bid according to the balanced bidding strategy, the bidders' behavior converges to the LEFNE. However, the validity of balanced bidding strategy has not yet been examined in practice.

In the existing literature on the VCG, the only concept that has been considered is the dominant strategy equilibrium, unlike that on the GSP, where many equilibrium concepts have been considered. However, when comparing the two mechanisms, it is worthwhile to use the same equilibrium concept. In the following, we list the results obtained from considering the LEFNE of the VCG.

Proposition 2. *The following statements on the VCG hold true.*

(i) The truth-telling equilibrium, i.e., the dominant strategy equilibrium, in the VCG is a LEFNE.

(ii) If bid profile b is a LEFNE in the VCG, every advertiser (except for the one submitting the highest bid) overbids or submits their true value.

(iii) In the truth-telling equilibrium of the VCG, the allocation of the ad-spots is efficient, as is the allocation of the ad-spots in all LEFNE.

(iv) The maximum revenue among LEFNE in the VCG is the same as that in the GSP.

(v) The maximum revenue among LEFNE of the VCG is achieved in the following LEFNE bid profile $\bar{b}^V$:

$$\bar{b}_k^V = v_{k-1} \text{ for each } k = 2, 3, ..., N \text{ and any } \bar{b}_1^V \text{ with } \bar{b}_1^V > \bar{b}_2^V.$$

(vi) The minimum revenue among LEFNE of the VCG is achieved in the following LEFNE bid profile $\underline{b}^V$:

$$\underline{b}_k^V = v_k \text{ for each } k = 2, 3, ..., N \text{ and any } \underline{b}_1^V \text{ with } \underline{b}_1^V > \underline{b}_2^V.$$

Proof. (i) Let b be a truth-telling bid profile. Take any $k > 1$. Then,

$$\alpha_k v_{d(k)} - p_k^V(b) - \left(\alpha_{k-1} v_{d(k)} - p_{k-1}^V(b)\right) = (\alpha_{k-1} - \alpha_k)(b_{d(k)} - v_{d(k)}). \tag{6}$$

Because $b_{d(k)} = v_{d(k)}$, the above equation must be zero. This means that b is a LEFNE.
(ii) Suppose b is a LEFNE. Then, (6) must be non-negative. This implies that for any $k > 1$, $b_{d(k)} \geq v_{d(k)}$.
(iii) It can be shown that the allocation is efficient for all LEFNE. Let b be a LEFNE. Assume that there exist i and j such that $v_i > v_j$ and $b_i \leq b_j$. From (ii) of this proposition, player i must make an overbid. Thus, $v_j < v_i \leq b_i \leq b_j$. In such a case, j's utility is improved by j switching their bid to $b'_j = v_j$ from b_j. This contradicts that b is a Nash equilibrium.
(iv) We know that the maximum revenue of the GSP is R_M defined in (4). We consider the maximum LEFNE revenue of the VCG. It can be easily verified that the maximum revenue in the LEFNE is attained by the bid profile b such that

$$b_k = v_{k-1}$$

for each $k > 1$ and $b_1 = 2v_1$. In this case, the revenue is

$$\sum_{k=1}^{K} p_k^V(b) = \sum_{k=1}^{K} \sum_{h=k+1}^{K} (\alpha_{h-1} - \alpha_h) b_h = \sum_{k=1}^{K} \sum_{h=k+1}^{K} (\alpha_{h-1} - \alpha_h) v_{h-1} = \sum_{k=1}^{K-1} k v_k (\alpha_k - \alpha_{k+1}) = R_M.$$

(v) and (vi) These are easily obtained from (i), (ii), and (iv) of this proposition. ■

From (ii) of Proposition 1 and (iii) of Proposition 2, both the GSP and the VCG mechanisms assure the efficient allocation of ad-spots under the assumption that a LEFNE bid profile is achieved. From (i) and (ii) of Proposition 2, it can be easily understood that the dominant strategy equilibrium in the VCG is the LEFNE with the least amount of revenue among all the LEFNE in the VCG. This questions the validity of Edelman et al. (2007) regarding the revenue comparison between the VCG and the GSP, because the dominant strategy equilibrium in the VCG is the LEFNE with the minimum revenue. Thus, if we adopt the LEFNE as the expected behavior of the bidders, we must compare both the minimum and maximum revenues between the GSP and the VCG. The fourth statement of Proposition 2 addresses this question by stating that both the maximum and minimum revenues of the two auction mechanisms are the same. Thus, when we use the LEFNE as the equilibrium concept in both mechanisms, it is unclear whether the GSP is more profitable than the VCG.

3 Experimental design

This section describes the experiment designed to compare two keyword auction mechanisms: the GSP and the VCG. The experiment consists of two treatments corresponding to each mechanism. The between-subject design was used, so that each subject participated in one treatment. Because the only difference between the two mechanisms is the payment rule, we controlled other factors in our experiment to eliminate other differences between the treatments.[8] Treatments and parameters are summarized in Table 1.

Treatments	VCG and GSP
Number of subjects per treatment	45
Number of groups per treatment	9
N	5
K	5
Values $(v_1, \ldots, v_5)$	$(180, 160, 120, 100, 90)$
CTRs $(\alpha_1, \ldots, \alpha_5)$	$(100, 80, 50, 40, 20)$

Table 1: Summary of the treatments and parameters of the experiment

The game played in the experiment is not the same as the game described in the previous section. In the previous section, the keyword auction is modeled as a one-shot game of complete information, but the keyword auction in practice is more like a dynamic game of incomplete information. As discussed in Edelman and Schwarz (2010), a well-modeled static game captures important points of the dynamic game while keeping the analysis tractable, and there is a long history in modeling the dynamic game of incomplete information as a one-shot game of complete information, such as the Bertrand model of oligopolistic competition. The theoretical analysis in the previous section followed these previous approaches. In the experiment, however, we were more interested in comparing the two mechanisms in a dynamic game of incomplete information, which is closer to a keyword auction in practice, rather than comparing the two under a one-shot game of complete information. The game played in the experiment is thus a dynamic game of incomplete information. If the theoretical results based on the one-shot game model hold in this experiment, then the results provide strong support for the theory and these approaches, although this design will not reject the theory even if the experimental results are not in line with them. In addition, knowing the rules of the game may be different from knowing how to play the game well. This experimental design provides more opportunities for the subjects to learn how to play the game in the course of the repeated play compared to the one-shot game of complete information. As the results in the next section show, the experimental results on revenues and efficiencies are in accordance with the theoretical results.

[8]Therefore, subjects were uninformed that bidding one's own value is the dominant strategy in the VCG. Although it is also interesting to see their behavior when they are informed about this feature of the VCG, this must be done in a different experiment.

The game played in the experiment is as follows. There are 5 players ($N = 5$) and 5 positions ($K = 5$). Subjects play the previously explained one-shot game of a keyword auction with one of the two mechanisms determined by the treatment for 100 rounds. Each round has the following steps: subjects simultaneously submit their bid to the auctioneer; the ad-spots are allocated in the descending order of the bids b, and ad-spots are allocated randomly in the case of ties; depending on the treatments, payment $p_k(b)$ of a subject in ad-spot k is determined as $p_k^V(b)$ or $p_k^G(b)$; and the profit of each i who obtain ad-spot k is given by the equation $u_i(b) = \alpha_k v_i - p_k(b)$. The subjects are paid for the sum of payoffs earned in all 100 rounds. The matching of the group is fixed across rounds. The CTRs $(\alpha_1, \ldots, \alpha_5)$ are commonly known to the subjects and are fixed for 100 rounds at $(100, 80, 50, 40, 20)$. The value of each subject is first determined from the interval of 50 to 250, and is fixed during the 100 rounds. They find out about their own value, but not about the value of others. The subjects are instructed about all of these settings. This game is essentially the same as the model of the dynamic game of incomplete information in Edelman and Schwarz (2010).

To add control to the comparison across treatments, we had only one value set, (180, 160, 120, 100, 90).[9] This is one limitation of this experiment. The number of groups is too limited for a comprehensive analysis with different values per group; therefore, we restricted the value set to one and conducted thorough analyses with this value set.[10] We chose values with two properties. First property is that v_1 and v_5 are not too close to the bounds of 50 and 250, respectively, in order to prevent the subjects from correctly inferring the rank of their value from the start. The more important property is that in both treatments, the LEFNE that yield the maximum and minimum revenue for the auctioneer are natural numbers. Thus $\bar{b}^G$, $\bar{b}^V$, $\underline{b}^G$, and $\underline{b}^V$ are all natural numbers. There are some existing theoretical and simulation studies that emphasize several distinct properties of $\underline{b}^G$. So we wanted these boundary bid profiles to be attainable in the experiment where the bids are restricted to whole numbers between 0 and 10000. With these experimental parameters, the boundary bids were as follows: $\underline{b}^V$ is $(161, 160, 120, 100, 90)$, $\underline{b}^G = (97, 96, 80, 56, 45)$, $\bar{b}^V = (181, 180, 160, 120, 100)$, and $\bar{b}^G = (117, 116, 100, 64, 50)$.[11]

3.1 Procedures

The details of the experimental procedures are as follows. First, after all subjects were seated at computers separated by partitions, they read slides with instructions at their own pace. Then, they answered the control questions. These answers were checked by the staffs, and the subjects continued with the questions until they correctly answered all of them. After every subject had answered correctly, the answers were reviewed by the experimenter to assure the public knowledge of the instructions.[12] Five practice rounds were run to familiarize the subjects with the computer program. Then, the subjects played 100 rounds of the keyword auction. Upon completion, they answered a short questionnaire while the payments were prepared. Both the experiment and the questionnaire were programmed and conducted with

[9] This feature of the experimental design, that there was only one value set, is not mentioned in the instructions to the subjects.

[10] Since the game in the experiment is a dynamic game of incomplete information with fixed value, the number of possible value sets for analyses equals the number of groups. For a keyword auction, we wanted to have a sufficient number of players N and positions K, and decided to have $N = K = 5$. The objective came at the cost of reducing of the number of groups in each treatment, but this is one salient point of the experimental design.

[11] The bid of the subject with the highest value, b_1, can be anything greater than b_2. Here, it is set to $b_1 = b_2 + 1$.

[12] By using slide instructions and allowing the subjects to read it at their own pace, we believe that they better understood the auction rules. Since the auction rules were difficult, the use of slides was necessary to insure their understanding, even at some cost of public knowledge of the instructions. We compensated for this lack of public knowledge by explaining the answers to the quiz aloud. We checked whether the subjects understood the auction rules by asking the subjects to rate the difficulty of understanding the instruction in the post-experiment questionnaire. We used a five-point scale from "Very Difficult (= 1)" to "Very Easy (= 5)." The questionnair results showed that the mode was "Very Easy" and the average score was 4.14. The instruction slides, as well as other materials, are available as supplementary materials at the journal's website.

the software z-Tree (Fischbacher, 2007). To control for the experimenter effect, all sessions were run by the same experimenter.

The practice rounds are designed to control for the subjects' learning. First, every subject had the same value of 150. To prevent them from learning about the others' behavior, each subject played with four computer-programmed players who bid randomly from the interval [0, 300]. Finally, to prevent different learning, random bids of the program were decided before-hand, so every subject faced the same random bids. The subjects knew they were playing a computer program during the practice rounds and with human subjects during the real 100 rounds.

Figure 1: Example of the decision-making screen

Let us describe the details of the subjects' decision-making procedure in each round. Before the first round of the auction, the subjects were informed about their value which remained constant throughout the 100 rounds. When the value was disclosed, the subjects had three minutes before the start of the first round. We did not suggest how to use these three minutes, although we intended it to be a thinking time for the subjects. In each round, the subjects were provided with the feedback information from the previous round and simultaneously submitted their bids. Figure 1 shows an example of the screen. To change their bids from the previous round, subjects had to submit their bid within 15 seconds. Otherwise, the same bid as in previous rounds was automatically submitted.[13] The feedback information was as follows: all the bids made by group members and their payment (in the descending order of bids), his or her own bid, the earned position, gain from the advertisement ($\alpha_k v_i$), payment, and payoff. In addition, because they had no time to take notes of the results, a graph of the history of bid profiles was provided. Each bid was plotted in black, and only their own bids were connected with a red line

[13]To ensure that our time limit was not too short, we asked the subjects to rate the time limit on a seven-point scale, and the number of times they were unable to submit a bid although they intended to. The number of subjects who answered that 15 seconds was too short (= 1 on the scale) was 2 (out of 90), the mode was neither long or short (= 4), and the mean was 3.58. The number of subjects who often (= 7 on the scale) intended to but were unable to submit a bid was 2 (out of 90). The mode of the frequency of bids not submitted was "a few" (= 5), and the average was 3.40. These results suggest that the time limit was not too short for the subjects.

across rounds. No information was lost by not connecting the others' bids, because the subjects did not know who acquired each position, but knew only which position was acquired with which size bid. The vertical axis of the graph has a range of 0 to 350. The range of the axis was decided by analogy to the range of value: 50 to 250.

This experiment was conducted in December, 2009 at Waseda University on 90 undergraduate students whom we recruited through university's web page with an ad for a part-time job. Each treatment had three sessions with 15 participants per session. Each session lasted for about two hours. At the end of the session, the subjects were paid in cash. The currency in the experiment was referred to as "points." The subjects' earnings were the sum of points earned in all 100 rounds exchanged at the rate of 400 points = 1 Yen. Subjects were also paid a participation fee of 1300 Yen. The average of the payments was 2282 Yen, which was about $25 at the time of the experiment.

4 Experimental results

The primary purpose of this experiment is to compare two major criteria of auction mechanisms: the auctioneer's revenues and the efficiencies of allocations. If the outcomes in both mechanisms are LEFNE, the maximum and minimum of the auctioneer's revenues are the same, and the allocations of ad-spots are efficient in both mechanisms.

The LEFNE is used as the theoretical benchmark, and the data is analyzed from this perspective. The rationale behind the use of the LEFNE (of the one-shot game of complete information) as the theoretical benchmark for the dynamic game of incomplete information is as follows: this concept has already been well analyzed and some properties are known; and for a bid to be LEF, it does not require the information about other's value (see Inequalities (2) and (3) in Section 2.2). One underlying consideration in analyzing the dynamic game of incomplete information as a one-shot game of complete information is that when a game is repeated for some time, participants will learn many characteristics of the other players. To take this into account, most of the statistical tests conducted in this study were based on the observations in the last 50 of the 100 rounds.

Another possible benchmark is to compare the results to the case where all the bids are made randomly. To see this, we computed the relative frequency of efficient outcomes and equilibrium outcomes with bids restricted to whole numbers in the range of 0 to 350.[14] If all the subjects choose their bids randomly from integers in the range of 0 to 350 with the same probability, the observed relative frequency in the experiment would be close to the computed relative frequencies. Because there are five players, the total number of outcomes in the computation is 351^5. The results of this computation are compared with the experimental results.

We first compare the revenues and efficiencies of the two mechanisms, and then examine the bid profiles to compare the frequency of the equilibrium outcomes.

4.1 Revenue

The auctioneer's revenue is the sum of payments by advertisers. Because the auctioneer decides which mechanism to use, a mechanism that yields higher revenues will be more attractive. The revenue will be higher in the GSP if the distributions of bids are the same in the two mechanisms; however, if the bid profiles are LEFNE, the range of revenues in the VCG and the GSP will coincide. In this experiment, the observed revenues were indifferent in the VCG and the GSP.

[14] The range of 0 to 350 was chosen arbitrarily. It is the same as the range shown in the graph of the decision-making screen (see Figure 1).

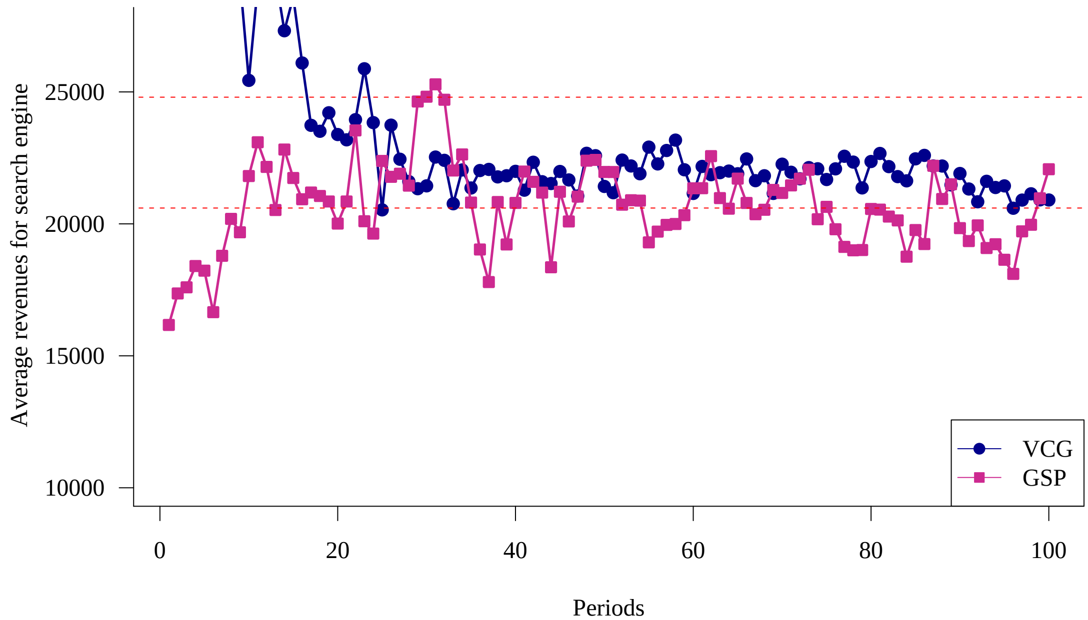

Figure 2: Comparison of the average revenues of the auctioneer across periods

First, we looked at the change in the revenues across periods. Figure 2 plots the average revenues over 100 periods by pooling data for each treatment. The two horizontal dotted lines are R_M and R_m. With the parameters used in the experiment, the lower bound R_m is 20600 and the upper bound R_M is 24800. From this graph, it is possible to see that after about round 50, the average revenue becomes relatively stable in both treatments. In the later rounds, the average revenues tend to lie in between the bounds of $[R_m, R_M]$. Also, the revenue tends to be closer to the lower bound R_m than the upper bound R_M. This is a noteworthy point, because the lower bound of the auctioneer's revenue has several distinguishing theoretical properties. The most notable one is that in the VCG, the lower bound revenue is obtained in the dominant strategy equilibrium. More properties of the lower bound are investigated in Cary et al. (2007), Bu et al. (2008), and Edelman and Schwarz (2010).

Summary statistics per group are listed in Table 2. The first two columns correspond to the within-group average and standard deviation of the auctioneer's revenues in the first half and the last half. In this experiment, there are nine independent groups per treatment. The top half corresponds to the groups in the VCG treatment, and the bottom half corresponds to the groups in the GSP treatment.

First, for each treatment, we looked at the auctioneer's revenue stability, measured by the standard deviation of the revenue across periods. To analyze whether the revenue became more stable with repetition, we compared the standard deviation of the auctioneer's revenue in the first half and the last half. In both treatments, all the groups' standard deviation of the revenues was greater in the first half than in the last half. Using the paired-sample sign test, we tested the null hypothesis that the median of the difference in the standard deviations equals zero. The null hypothesis was rejected at the 5% level (p-value = 0.004). To compare the stability across treatments, we used the Wilcoxon rank sum test to compare the distribution of the groups' standard deviation of revenues in the last half. The standard deviation was lower in the VCG than in the GSP. The difference in the distribution was statistically significant (p-value = 0.019).

Next, although the revenue in the VCG seems to be larger than in the GSP, the difference between the auctioneer's revenues in the two treatments is not statistically significant in the last half. By using the Wilcoxon rank sum test to compare the distribution of the groups' average revenue in the last half across treatments, the p-value was found to be 0.062.[15]

Finally, we analyzed whether the median of the average auctioneer's revenue per group was in the bounds of LEFNE. The 95% confidence interval of the median obtained from the acceptance region of the sign test (Gibbons and Chakraborti, 2010) is [18736.8, 22059.6] for the GSP and [20278.2, 22621.2] for the VCG. Although the lower bound of the confidence interval is lower than the minimum revenue of the LEFNE, using the sign test, we could not reject the null hypothesis that the median is equal to the minimum revenue. In sum, the experimental results for the auctioneer's revenue agree with the theoretical benchmark.

Observation 1. The auctioneer's revenue was more stable in the VCG than in the GSP. The difference in the auctioneer's revenue between the two treatments was not statistically significant. Also, in both treatments, the median of the average auctioneer's revenue lies close to the bounds of $[R_m, R_M]$. The revenue was closer to the lower bound than the upper bound in both treatments.

[15]The result is significant at the 10% level. However, since we have only one value set, we wanted to be extremely careful when rejecting the null hypothesis. Therefore, we state that the revenue in the VCG is not statistically different from that in the GSP.

	Average revenue (standard deviation)		Average efficiency (standard deviation)		# of N.E. bid profiles		# of LEFNE bid profiles		Average # of best responding bids		Average # of LEF bids	
	First half	Last half	First half	Last half	First half	Last half	First half	Last half	First half	Last half	First half	Last half
VCG												
Group 1	18064.6	20278.2	0.73	0.92	1	16	0	2	2.76	3.90	1.84	3.06
	(4495.70)	(1915.41)	(0.22)	(0.12)								
Group 2	25432.2	25132.0	0.70	0.85	0	2	0	0	2.26	2.84	1.80	2.32
	(6475.92)	(3645.33)	(0.24)	(0.17)								
Group 3	25214.0	19955.2	0.79	0.87	3	2	0	0	2.54	3.30	1.98	2.28
	(6758.6)	(801.77)	(0.22)	(0.13)								
Group 4	20837.6	22107.6	0.75	0.89	0	8	0	2	2.42	3.34	1.76	2.80
	(3406.26)	(1304.83)	(0.18)	(0.11)								
Group 5	24781.0	22329.6	0.82	0.93	3	14	0	3	2.64	3.78	2.16	3.02
	(4359.94)	(2012.54)	(0.24)	(0.09)								
Group 6	31744.6	21778.8	0.81	0.94	1	9	0	1	2.54	3.50	1.96	2.46
	(30422.79)	(1772.32)	(0.19)	(0.10)								
Group 7	34890.6	21167.4	0.72	0.92	2	0	0	0	2.18	2.76	1.80	2.38
	(25780.26)	(915.86)	(0.22)	(0.09)								
Group 8	24212.4	22621.2	0.69	0.91	0	4	0	2	2.14	3.32	1.68	2.48
	(6311.68)	(1873.26)	(0.24)	(0.09)								
Group 9	23739.4	21490.8	0.77	0.94	0	1	0	1	2.30	3.26	2.08	2.96
	(4176.47)	(1420.14)	(0.21)	(0.08)								
All	25435.2	21873.4	0.75	0.91	10	56	0	11	2.42	3.33	1.90	2.64
	(14772.57)	(2371.89)	(0.22)	0.12								
GSP												
Group 1	21329.4	22059.6	0.83	0.98	1	0	0	0	2.66	3.12	1.96	2.36
	(3675.64)	(1029.65)	(0.18)	(0.04)								
Group 2	19258.0	20829.6	0.75	0.88	1	1	0	0	2.56	2.66	1.68	2.04
	(3574.42)	(3308.50)	(0.21)	(0.15)								
Group 3	13459.0	17727.4	0.76	0.86	0	0	0	0	2.14	2.70	1.40	1.92
	(6617.31)	(4605.58)	(0.16)	(0.13)								
Group 4	22671.4	18736.8	0.79	0.87	0	0	0	0	2.52	2.74	1.84	2.10
	(5269.26)	(3358.58)	(0.23)	(0.11)								
Group 5	27541.2	23712.0	0.64	0.88	0	2	0	0	1.74	2.86	1.70	2.38
	(3578.95)	(2567.63)	(0.25)	(0.12)								
Group 6	16751.2	20187.0	0.76	0.88	0	1	0	0	2.06	2.56	1.62	2.10
	(3541.86)	(2097.21)	(0.21)	(0.14)								
Group 7	23376.6	19063.2	0.75	0.80	0	0	0	0	2.22	2.56	1.96	1.78
	(6318.98)	(4930.46)	(0.19)	(0.14)								
Group 8	20463.0	20201.0	0.78	0.89	0	5	0	0	2.54	3.14	1.84	2.08
	(4251.03)	(1988.39)	(0.15)	(0.10)								
Group 9	23376.0	21125.2	0.84	0.93	0	1	0	0	2.44	2.98	1.74	1.90
	(5912.53)	(3394.94)	(0.17)	(0.10)								
All	20914.0	20404.6	0.77	0.88	2	10	0	0	2.32	2.81	1.75	2.07
	(6216.12)	(3649.25)	(0.20)	(0.13)								

Table 2: Per group summary statistics

4.2 Efficiency

If bid profiles are LEFNE in both mechanisms, the allocations are efficient. However, as we will see in the next subsection, the bid profiles observed in the experiment were rarely LEFNE. This does not imply that most allocations were inefficient, because being a LEFNE is only a sufficient condition for a bid profile to be efficient. Therefore, it is meaningful to compare the efficiencies of allocations in the two mechanisms. In keyword auctions, the allocations are efficient if the ad-spots are assortative, i.e., an advertiser with the k-th highest value obtains the k-th position. Our main analysis uses surplus as the measure of efficiency, which is the ratio of the total surplus of the allocation to the maximum possible surplus, normalized by the minimum possible surplus. In addition to the surplus, the Kendall-τ correlation is used in the analysis to measure efficiency. This is to take into account the fact that, for efficiency, the important factor is not the absolute size of the bid but the ranking of bids with respect to the value. Using both measures, no significant difference between the efficiencies of the GSP and the VCG was found.

First, let us compare the frequency of efficient allocations to the benchmark of random bidding.[16] According to the computation, the relative frequency of efficient allocations is 0.00857 ($= 456746106967/351^5$). The relative frequency of efficient allocations observed in the experiment was 0.09 in the GSP and 0.112 in the VCG, which is much higher than the benchmark. To compare in greater detail, for each treatment, we calculated the relative frequency of efficient allocations per group. Using the sign test, we tested the null hypothesis that the median of the relative frequency of the efficient allocation per group equals 0.00857. The null hypothesis was rejected at the 5% level in both treatments with a p-value of 0.004.

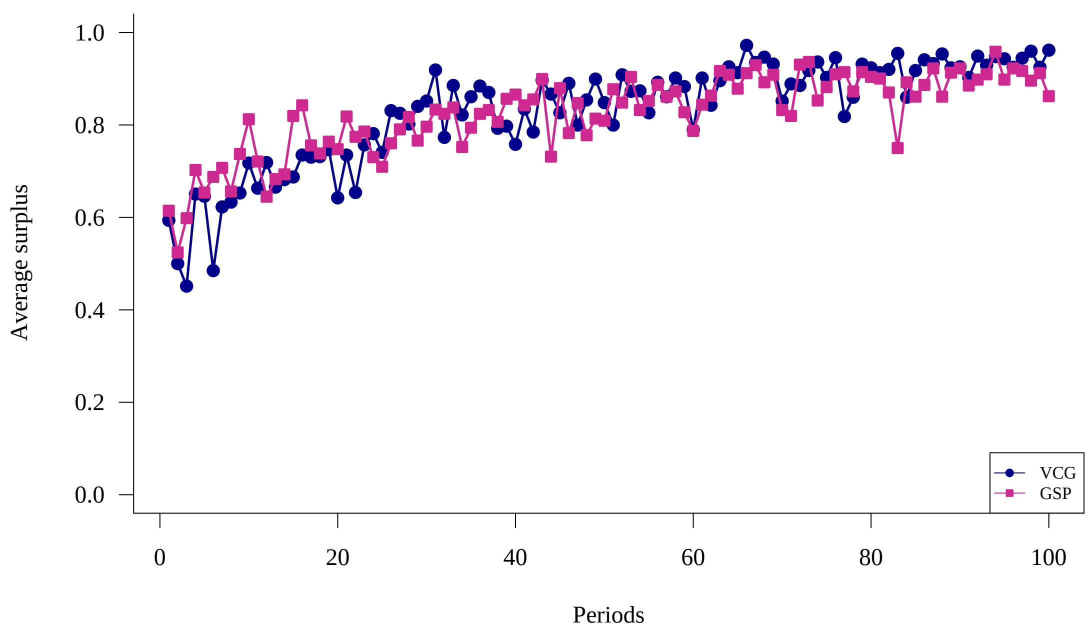

Figure 3: Comparison of the average surplus across periods

[16]In the computation, we broke the ties so as to maximize the number of efficient allocations. Thus, if the bids are $b_1 \geq b_2 \geq \cdots \geq b_5$, we counted it as being efficient. Another possibility is to count it as inefficient when some equality holds in the equation, which would give the minimum number of efficient allocations. The important point is that although we counted the maximum possible number of efficient allocations, the outcomes of the experiment were much better in terms of the relative frequency of efficient outcomes.

Figure 3 plots the average surplus for each period. We measured the surplus by

$$\frac{\sum_{k=1}^{K} \alpha_k v_{d(k)} - \sum_{k=1}^{K} \alpha_k v_{K-k+1}}{\sum_{k=1}^{K} \alpha_k v_k - \sum_{k=1}^{K} \alpha_k v_{K-k+1}},$$

which is the ratio of the total surplus to the maximum possible surplus–obtained when the allocation of ad-spots is assortative–normalized by the minimum possible surplus–obtained when the allocation of ad-spots is reversed, i.e., $b_5 > \cdots > b_1$. This definition of the surplus is along the same lines as that of Kagel and Levin (2001), which is the ratio of the total surplus of allocations to the highest possible surplus. We have normalized that ratio by the minimum possible surplus. Figure 3 reveals that in the GSP and the VCG, the relative efficiencies improved with repetition. The average surplus of the first half and the last half is listed in the third and fourth columns, respectively, of Table 2 (with the standard deviations in parentheses). For both treatments and for all the groups, the average surplus of the first half was lower than that of the last half. Using the sign test for paired samples, the null hypothesis that the median of the within-group difference in the average surplus of the first half and the last half equals zero was rejected with a p-value of 0.004 for both treatments. Also, according to the figure, there was no difference in the efficiency of the two mechanisms. Treating each group's average as a sample and using the Wilcoxon rank sum test, we tested the null hypothesis that the two treatments have the same distribution of surplus. The p-value was 0.258. We conducted the same test using the Kendall-τ correlation, obtaining an insignificant result with a p-value of 0.508.

Observation 2. The frequency of efficient allocations observed in the experiment was much higher than the benchmark of random bidding. There was no difference in the efficiency of the GSP and the VCG. In both treatments, the average surplus increased with repetition.

4.3 Equilibrium

One benefit of using experimental methods is that we can specify the value. This allows for a direct test of whether a bid profile is in equilibrium. In the benchmark of random bidding, the relative frequency of Nash equilibrium bid profiles is 77.316×10^{-5} $(= 4119147603/351^5)$ in the VCG and 2.501×10^{-5} $(= 133265484/351^5)$ in the GSP.[17] The relative frequency of the LEFNE bid profiles is 675.707×10^{-8} $(= 35999271/351^5)$ in the VCG and 7.579×10^{-8} $(= 403760/351^5)$ in the GSP. These results illustrate two features of the auctions. First, if all bids are determined randomly, the probability of observing bid profiles that are either Nash equilibria or LEFNE is extremely low. Second, there are more Nash equilibrium and LEFNE bid profiles in the VCG than in the GSP. Nash equilibrium bid profiles are 30 times more likely and LEFNE bid profiles are 90 times more likely in the VCG than in the GSP. In this experiment, there were very few Nash equilibrium bid profiles, and LEFNE bid profiles were rarely observed. Equilibrium bid profiles were more frequent in the VCG than in the GSP.

The top two graphs in Figure 4 illustrate the number of bid profiles that were a Nash equilibrium and a LEFNE in each period. The number of bid profiles in each period was nine, so half of the bid profiles were not even a Nash equilibrium. There were no LEFNE bid profiles in the GSP; there were 11 LEFNE bid profiles in the VCG. As apparent from the figure, equilibrium bid profiles were more frequent in the VCG. We calculated the occurrence of Nash equilibrium and LEFNE outcomes per group in the first and the last 50 rounds for each treatment (listed in the fifth through eighth columns of Table 2). Using the Wilcoxon rank sum test, the tested null hypothesis was that there is no difference

[17] As with the calculation of efficiency, we counted the allocations in such a manner that the number of equilibrium outcomes is maximized (see Footnote 16). The relative frequency of the experiment is much higher than the relative frequency of the computation, although we counted the equilibrium in this manner. We also computed the ratios for the case where the number of equilibrium outcomes would be minimum. The ratios of equilibrium outcomes in the VCG to the GSP did not differ greatly.

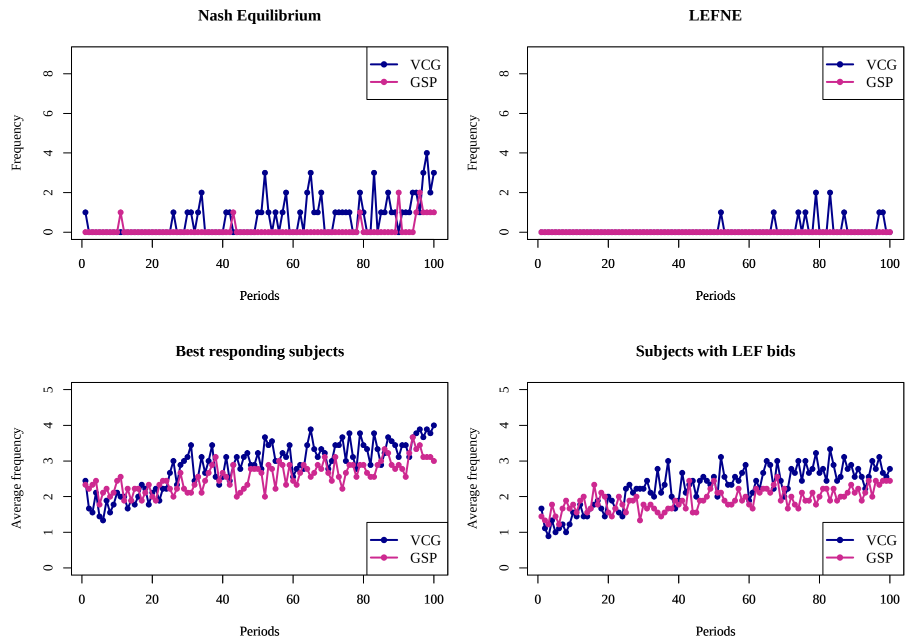

Figure 4: Comparison of the occurrence of equilibrium bid profiles

between the two treatments in the distributions of the number of equilibrium outcomes in the last 50 rounds. The p-values were 0.005 for LEFNE and 0.021 for Nash equilibrium outcomes.

The bottom two graphs in Figure 4 compare across treatments the average number of subjects in a group whose bid is the best response to the others' bids (right) and whose bid is LEF (left). Clearly, if the graph reaches 5, all the groups' outcomes would be a Nash equilibrium or a LEFNE. As the figures depict, both graphs tend to increase over periods for both treatments. Also, the graph is higher for the VCG than for the GSP. To see the results in further detail, the per-group averages of the first half and the last half are listed in the ninth to eleventh columns in Table 2. Comparing the top and bottom half rows in column 10, we can reject the null hypothesis that the distribution is the same in the GSP and the VCG at the 5% level (Wilcoxon rank sum test, p-value = 0.005). We compared the last column similarly, and the null hypothesis was rejected (Wilcoxon rank sum test, p-value = 0.002).

Observation 3. The observed number of equilibrium outcomes was higher in the VCG than in the GSP. Also, the per-group averages of the number of best responding subjects and the number of LEF bids were higher in the VCG than in the GSP.

Let us end this subsection with a discussion on the possible reasons for there being more equilibrium outcomes in the VCG than in the GSP. First, it is clear from the computation results that the number of Nash equilibria and LEFNE is larger in the VCG than in the GSP. Although this was the result of computations for a specific environment, the difference in the number of equilibria is noteworthy. The VCG has more equilibria than the GSP, and this difference could have accounted for the experimental observation. Another possibility is the difference in the existence of truth-telling equilibrium. Mizukami et al. (2009) argued that if it is an equilibrium for all to state truthfully in a mechanism with multiple equilibria, this truth-telling equilibrium may become the focal point among all the equilibria. They showed that this was indeed the case by comparing two simple mechanisms with and without a truth-telling equilibrium in an experiment, and observing that the Nash equilibrium was observed much more frequently in the mechanism with the truth-telling equilibrium. In keyword auctions, it is a LEFNE for all to bid their value in the VCG, whereas such a bid profile is not always a Nash equilibrium in the GSP. If the truth-telling equilibrium were a focal point, the revenue would be lower in the VCG than in the GSP. Also, because the GSP lacks an equilibrium that is a focal point, it may have been more difficult for the subjects to coordinate on the equilibrium, lowering the number of equilibrium bid profiles in the GSP compared to that in the VCG. Finally, given others' bids, it is easier to find best response in the VCG than in the GSP in the following sense: the payoff maximizing position will be unique in most cases in the VCG but not in the GSP. In the VCG, player i's best response given the others' bids b_{-i} is to bid in the interval of[18] $[max\{b_j : j \neq i, b_j \leq v_i\}, min\{b_j : j \neq i, b_j \geq v_i\}]$. Therefore, as long as the endpoints of the interval are not equal to the value, the payoff maximizing positions will be unique. In the GSP, however, the best response may not always form a single interval, but may consist of several different intervals. For example, consider the case where the others' bids are $b_{-i} = (100, 78, 60, 55)$ and $v_i = 100$. The payoffs from obtaining position 1 to 5 by bidding accordingly will be 0, 1760, 2000, 1800, and 2000. The payoff maximizing positions of player i are positions 3 and 5, and the corresponding range of best response bids consist of two intervals [0, 55] and [60, 78]. Of course, the ease of best responding to the bid of others does not imply the ease of being in equilibrium, but these properties may cause the difference in the equilibrium outcome in the two mechanisms.

4.4 Bidding behavior

Unlike other sections where the performance of the mechanisms was compared, this section investigates the bidding behaviors of the subjects. First, we examine the distribution of bids in each treatment with respect to $\bar{b}$ and $\underline{b}$. The bids in both mechanisms tend to converge to a range $[\bar{b}, \underline{b}]$. Like the observations in the sealed-bid second-price auction experiments, there is a tendency for the subjects to overbid in the VCG, although this tendency is weak and the median is close to the value. Also, we look at the subjects' responses to the past bids of others. Although the Nash equilibrium bid profiles were rarely observed in the experiment, given the bids of others in the previous period, the subjects' bids were weakly improving their payoffs in most cases in both mechanisms.

Table 3 lists the summary statistics of the bidding behavior that we will use throughout this section. The statistics are calculated conditional on treatments and values. The first three rows list the bids' average, median, and standard deviation of the last half. From the average and median, it is clear that the bids are higher in the VCG than in the GSP. There is no general tendency in the standard deviations of the two mechanisms.

Another interesting question is whether the subjects' most experienced position coincides with the rank of their value. The number of subjects whose most experienced position equals the ranking of their value is listed in the fourth row of Table 3. It was almost the same in the GSP and the VCG.

Figure 5 shows 10 graphs that depict, conditional on treatments and values, the median bids and 95% confidence interval calculated from the acceptance region of the sign test. The thick dark blue line shows the median, and the light blue regions show the 95% confidence interval. The dotted horizontal lines are $\bar{b}$ and $\underline{b}$. The top row shows graphs for the VCG and the bottom row shows graphs for the GSP. Each column corresponds to the values $v_1, \ldots, v_5$.

From the graphs for the VCG treatment, one can see that the subjects with lower values tend to overbid, and subjects with higher values tend to underbid. This finding of overbidding was similar to that of the sealed-bid second-price auction experiments with a single unit of good (Kagel et al., 1987; Kagel and Levin, 1993; Harstad, 2000; Cooper and Fang, 2008) and multiple units of good (Kagel and

[18] If the lower bound is empty, it should be set to zero. If the upper bound is empty, it can be arbitrarily set to any number greater than or equal to the value.

	VCG						GSP					
	180	160	120	100	90	All	180	160	120	100	90	All
Bids (with corrections for outliers)*												
Average (last half)	162.31	150.46	128.04	109.62	105.11	131.11	112.29	101.80	71.77	55.47	50.57	78.26
Median (last half)	159.00	150.50	124.00	105.00	100.00	121.00	101.00	100.00	75.00	55.00	55.00	78.00
Standard deviation (last half)	52.09	29.36	18.44	14.65	19.47	37.46	32.92	24.83	27.79	19.11	27.26	36.39
Number of subjects whose most experienced position equals to the ranking of value	6	3	8	3	6	26	6	3	5	5	5	24
Average number of myopic improving bids per subject (out of 99 bids)	84.33	80.33	82.00	85.78	86.89	83.87	89.33	87.44	88.89	84.11	83.22	86.60
Average number of myopic best responding bids per subject (out of 99 bids)	58.00	52.44	47.89	54.56	50.00	52.58	63.33	57.00	49.11	64.67	49.56	56.73
Average number of best response to the average of the past bids (out of 99 bids)	50.78	50.56	58.44	64.89	58.33	56.60	45.00	50.78	22.78	54.67	33.89	41.42

* In the experiment, only a few extreme overbiddings were observed. We exclude them merely for the statistics in these three rows, because the descriptive statistics better describe the overall tendency without them. For each value, we excluded the bids that were three standard deviations away from the mean. The number of observation excluded from the analysis for values $v_1, \ldots, v_5$ were 2, 6, 7, 6, 1 for the VCG and 2, 12, 3, 3, 0 for the GSP, respectively.

Table 3: Summary statistics of the bidding behavior per treatment and value

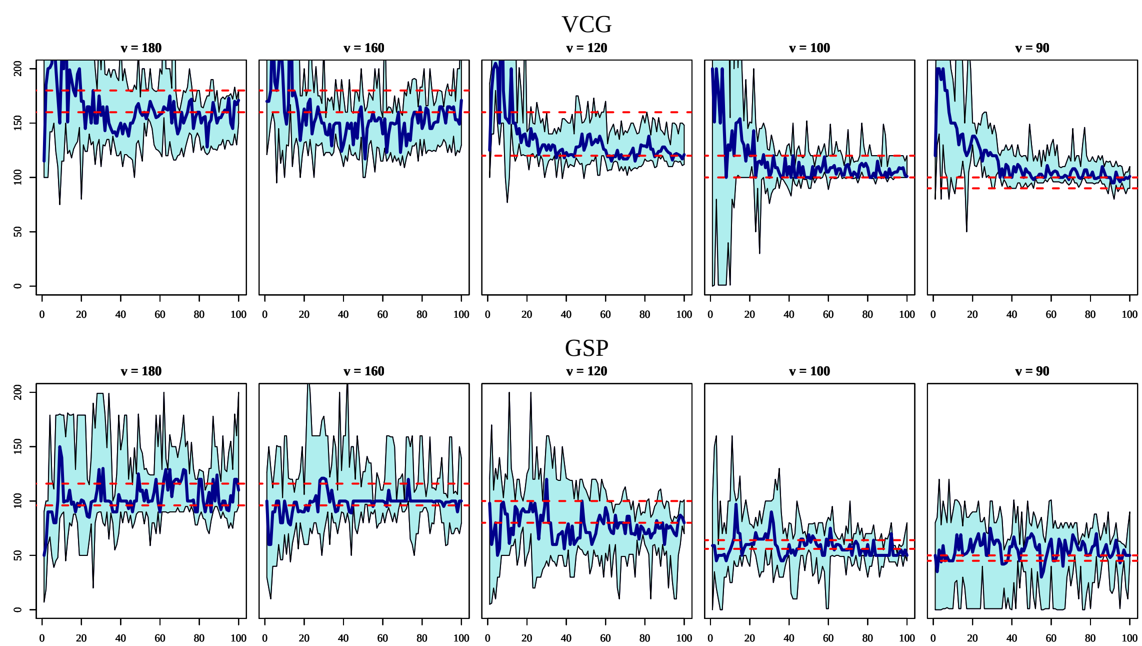

Figure 5: Median bid and 95% confidence intervals in each period for each treatment and each value (The top row shows graphs for the VCG and bottom row shows graphs for the GSP. Dotted horizontal lines are $\bar{b}$ and $\underline{b}$.)

Levin, 2001.)[19] However, the degree of overbidding is not large, as the bids tend to lie in between their own value and the value immediately above, which corresponds to $\underline{b}^V$ and $\bar{b}^V$, respectively. This tendency that the bids are in the bounds of LEFNE is also observed in the GSP. In most graphs of both treatments, the bids tend to converge to a range within $\bar{b}$ and $\underline{b}$. Thus, although the bid profiles are not in equilibrium, subjects' bids tend to converge toward a bid profile that is in the range of bids that give the upper and lower bounds to the auctioneer's revenue among the LEFNE bid profiles.

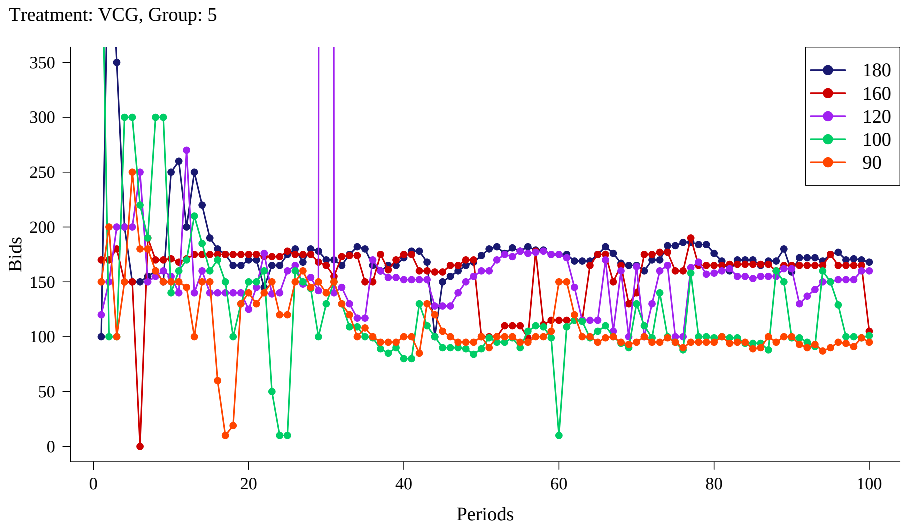

Figure 6: Example of bid transition in group (Treatment = VCG; Group = 5)

The exception to the above tendency is the bids made by subjects with lowest value. To investigate this behavior, we looked into the bid transition of each group. Figure 6 shows an example of bid transition in the VCG. In this group, bids of subjects with the lowest and second lowest value drift in parallel with each other, and the other three drift in parallel with each other. This tendency that some bids drift along is apparent in most of the groups, and it is most common for bids of subjects with values v_5 and v_4. The graphs of all groups are available as supplementary material. One possible interpretation of this bidding behavior is that some proportion of the population has a spiteful preference. By spiteful preference, we mean people having preference over not only their own revenue but also others' payments: for the same revenue, such people prefer to raise the payments of others. In keyword auctions, this simply means that people prefer to raise their bid as long as they obtain the same position. In the post-experiment questionnaire, we asked whether they increased their bid in order to increase the payments of people in higher positions. The answer was chosen from strongly disagree (=1) to strongly agree (=7) in a seven-point scale. The average answer was 4.76 in the VCG and 4.20 in the GSP, and "agree" was the mode for both treatments. If the subjects with the lowest value have this preference in the VCG, then we can expect the bids of subjects with v_4 and v_5 to drift in parallel with each other.[20]

[19] Chen and Takeuchi (2010) found that there is no tendency to overbid in the multi-object package auction using the VCG mechanism.

[20] For example, consider the bid of subject i with the lowest value v_5 in the VCG. Without this preference considered, subject i would be indifferent to bidding any number between 0 and $b_{d(4)}$. With this preference, i would prefer to bid $b_{d(4)} - \varepsilon$. Now, if subject j with the next lowest value v_4 is faced with subject i with this preference, j would not prefer to raise the bid to $b_{d(3)} - \varepsilon$, even if he or she had the spiteful preference. If j do have a spiteful preference, he or she could expect bidder i to raise their bid even more, lowering j's revenue in the following period. In another question in the post-experiment questionnaire, we asked whether they decreased their bid in order to prevent the subject in the lower position from raising their bid. The

Finally, we analyzed how the subjects were adjusting their bids to the past bids of others. First, we checked whether subjects' bids improved their payoffs, with respect to the others' bids from the previous period. If the subjects were making their decisions thoughtfully, it is reasonable to believe that the subjects based their decisions on the past bids of others. This analysis is in the same spirit as Cary et al. (2007), where they analyzed, both theoretically and using simulations, the convergence of bidding behavior in repeated auctions where the bidders were assumed to predict that others would behave in the next period as they did in the current period. To check the tendency of subjects' bids, we introduced two new concepts, *myopic improvement* and *myopic best response*. Let $b_{i,t}$ be the bid made by subject i in period t and $b_{-i,t}$ be the bids of subjects other than i in period t. We say that subject i's bid in period $t+1$ is a *myopic improvement* if $u_i(b_{i,t+1}, b_{-i,t}) - u_i(b_{i,t}, b_{-i,t}) \geq 0$, and *myopic best response* if $b_{i,t+1} \in \text{argmax}_{b'_{i,t}}\ u_i(b'_{i,t}, b_{-i,t})$.[21] One might think that individuals would consider more deeply into the past than just one previous period. For example, an individual may best respond to the average of all the previous bids by others. Therefore, we calculated the number of bids $b_{i,t+1}$ such that $b_{i,t+1} \in \text{argmax}_{b'_{i,t}}\ u_i(b'_{i,t}, \hat{b}_{-i,t})$ where $\hat{b}_{-i,t}$ is a vector of average bids by others in all previous periods ranked in the order of bids. Elements of the vector are the average of the highest bids, second highest bids, third highest bids, and lowest bids of others in each period.

The average frequency of myopic improvement bids per subject was 83.87 in the VCG and 86.6 in the GSP (Table 3). The average per subject frequency of myopic best response bids was 52.58 in the VCG and 56.73 in the GSP. Using the Wilcoxon rank sum test, the null hypothesis that the distribution of the per subject frequency of myopic best response bids is the same was almost rejected at the 5% level with the p-value of 0.051. This result is surprising since, as argued in the previous section, choosing the best response may be more difficult in the GSP because the bid range that is the best response to the others' bids may consist of more than one interval. Finally, contrary to the myopic best response, the per subject frequency of the best response bids to the average of all the past bids of others is higher in the VCG. The average was 56.60 in the VCG and 41.42 in the GSP. This difference was also statistically significant. Using the Wilcoxon rank sum test, the null hypothesis was rejected at the 5% level with the p-value of 0.002.

Observation 4. The median of the bids tends to lie in the bounds of $\bar{b}$ and $\underline{b}$. There was some overbidding in the VCG among subjects with lower bids, but this tendency was weak. The number of myopic improvement bids was high in both mechanisms. The subjects in the GSP tend to best respond myopically compared to the VCG, whereas subjects in the VCG tend to best respond to the average of all the past bids.

5 Conclusion

This paper compared two mechanisms for keyword auctions: the GSP and the VCG. The GSP is being used in real keyword auctions, whereas the VCG, where truth-telling is the dominant strategy, is not.

This paper makes two contributions to the existing keyword auction literature. First, theoretically, the properties of the LEFNE in the VCG were investigated. We proved that the ad-spot allocations are efficient in any LEFNE in the VCG, and that the range of the revenue obtained by the auctioneer in the

mean was 4.36 in the VCG and 4.31 in the GSP with the mode "agree" in both treatments. We can expect the bid of i and j to drift at some bid. A reasonable candidate for this bid is $\bar{b}_5^V = \underline{b}_4^V$, the threshold bid where j will prefer to obtain position 5 instead of 4 for any $b_{d(5)}$ higher than this number. If the probability of a subject being spiteful is above 0.5, we can expect the median of b_5 to drift at or just below $\bar{b}_5^V$, but the median of b_4 to drift at or just above $\underline{b}_4^V$.

[21] In line with Cary et al. (2007), if the subjects are following the "greedy bidding strategy," their bids will always be the myopic best response. In their analysis, they also specify a particular bidding behavior called "balanced bidding." They theoretically show that when there are more than three ad-slots and all players change their bids simultaneously, even if the bids are 100% myopic best response, the bids may not converge to equilibrium. See also Footnote 7.

LEFNE in the GSP coincides with that in the VCG. Thus, not only do the lower bounds of the revenues coincide in the GSP and the VCG as shown in Edelman et al. (2007), but the upper bounds of the revenues also coincide in the two mechanisms.

Second, this paper analyzed the results of the first laboratory experiment conducted to compare the VCG and the GSP. The main observations of the experiment are summarized as follows:

- The differences between the revenues of the two mechanisms were insignificant. In both mechanisms, the average revenues were approaching the lower bound of revenues among the LEFNE bid profiles.
- Efficiencies of the allocations improved with repetition in both mechanisms, and differences between the efficiencies of the allocation were indifferent in the two mechanisms.
- The observed number of equilibrium bid profiles was higher in the VCG than in the GSP.
- The subjects' bids were in the bounds of the LEFNE. Also, subjects' bidding behaviors showed some learning as their bids weakly improved their payoff with respect to others' bids in the previous period.

Theoretical results proven in Section 2 suggest that when the LEFNE is used as an equilibrium concept for analysis, the revenue and efficiency are expected to be similar in the GSP and the VCG. The results of the experiment are in line with the theoretical prediction. Although the bid profiles observed in the experiment were rarely LEFNE, the auctioneer's revenues were in the bounds of LEFNE and the efficiencies were high in both mechanisms. Therefore, when only the revenue and efficiency are under consideration, both mechanisms yield similar results. However, when the easiness to achieve equilibrium is considered, the VCG has some advantage over the GSP. First, there is a dominant strategy in the VCG. Second, given others' bids, the range of best response bids will be a unique interval in the VCG but might consists of many intervals in the GSP. Also, the number of possible equilibrium outcomes is larger in the VCG than in the GSP. Finally, in the experiment, the number of occurrences of the Nash equilibrium and LEFNE was higher in the VCG than in the GSP.

References

[1] Aggarwal, G., Goel, A., Motvani, R., 2006A. Truthful auctions for pricing search keywords. Proceedings of the 2006 ACM Conference on Electronic Commerce, 1–7. Ann Arbor, MI, 2006.

[2] Aggarwal, G., Muthukrishnan, S., Feldman, J., 2006B. Bidding to the top: VCG and equilibria of position-based. Workshop on Online and Approximation Algorithms (WAOA06).

[3] Bu, T.-M., Deng, X., Qi, Q., 2008. Forward looking Nash equilibrium for keyword auction. Information Processing Letters 105(2), 41–46.

[4] Cary, M., Das, A., Edelman, B., Giotis, I., Heimerl, K., Karlin, A. R., Mathieu, C., Schwarz M., 2007. Greedy bidding strategies for keyword auctions. Proceedings of the 8th ACM Conference on Electronic Commerce, 262–271. San Diego, California, 2007.

[5] Chen, Y., Takeuchi, K., 2010. Multi-object auctions with package bidding: An experimental comparison of Vickrey and iBEA. Games Econ. Behav. 68 (2), 557–579.

[6] Cooper, D., Fang, H., 2008. Understanding overbidding in second price auctions: An experimental study. Econ. J. 118, 1572–1595.

[7] Edelman, B., Ostrovsky M., 2007. Strategic bidder behavior in sponsored search auctions. Decision Support Systems 43(1), 192–198.

[8] Edelman, B., Ostrovsky M., Schwarz M., 2007. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. Amer. Econ. Rev. 100(2), 597–602.

[9] Edelman, B., Schwarz M., 2010. Optimal auction design and equilibrium selection in sponsored search auctions. Amer. Econ. Rev. 97(1), 242–259.

[10] Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Exper. Econ. 10(2), 171–178.

[11] Gibbons, J., Chakraborti, S., 2010. Nonparametric statistical inference. NewYork: Taylor & Francis.

[12] Harstad, R., 2000. Dominant strategy adoption and bidders experience with pricing rules. Exper. Econ. 3(3), 261–80.

[13] Kagel, J., 1995, Auctions: A survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, New Jersey/ West Sussex, pp. 501–585.

[14] Kagel, J., Harstad, R., Levin D., 1987. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. Econometrica 55, 1275–1304.

[15] Kagel, J., Levin, D., 1993. Independent private value auctions: bidder behavior in first-, second-, and third-price auctions with varying number of bidders. Econ. J. 103, 868–879.

[16] Kagel, J., Levin, D., 2001. Behavior in multi-unit demand auction: Experiments with uniform price and dynamic Vickrey auctions. Econometrica 69(2), 412–454.

[17] Kagel, J., Levin D., 2008. Auctions: A Survey of experimental research, 1995–2008. Working Paper.

[18] Lahaie, S., Pennock, D., Saberi A., Vohra R., 2007. Sponsored search auctions. In Nisan N., Roughgarden T., Tardos E., Vazirani V. (Eds.), Algorithmic Game Theory. Cambridge University Press, Cambridge, pp. 699–716.

[19] Mizukami, H., Nihonsugi T., Wakayama T., 2009. Is the truth-telling equilibrium salient?: An experiment on direct Nash implementation. mimeo.

[20] Ostrovsky M., Schwarz M., 2009. Reserve prices in internet advertising auctions: A field experiment. Research Papers 2054, Stanford University, Graduate School of Business.

[21] Smith, V., 2002. Method in experiment: Rhetoric and reality. Exper. Econ. 5, 91–110.

[22] Varian, H., 2007. Position auctions. Int. J. Ind. Organ. 25, 1163–1178.

[23] Varian, H., 2009. Online ad auctions. Amer. Econ. Rev. 99(2), 430–34.