Economic and Subjective Measures of the Perceived Value of Aesthetics and Usability

TAMAR BEN-BASSAT, JOACHIM MEYER, and NOAM TRACTINSKY
Ben Gurion University of the Negev

The assessment of the relative value of different design features for users is of great interest for software designers. Users' evaluations are generally measured through questionnaires. We suggest that other evaluation methods, including economic measures, may provide different estimates of the relative value of features. In a laboratory experiment we created four versions of a data-entry application by independently manipulating the system's usability and aesthetics. Users' evaluations of the four experimental systems were obtained in a within-subjects design. In addition, five between-subjects experimental conditions were created, based on the evaluation method (questionnaire alone or auction and questionnaire), monetary incentives (present or absent), and experience in using the system (present or absent). In questionnaire-based responses, the systems' usability affected evaluations of usability as well as aesthetics. Similarly, the systems' aesthetics affected evaluations of both aesthetics and usability. Questionnaire-based evaluations of usability and aesthetics were not affected by experience with the system or by monetary performance incentives. Auction bids were only influenced by the system's usability: bids corresponded to the objective performance levels that could be attained with the different systems. The results suggest that by using economic methods, researchers and practitioners can obtain system evaluations that are strongly related to performance criteria and that may be more valid when the evaluation context favors task-oriented performance.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Information**]: User Interfaces—*Evaluation/methodology, screen design, theory and methods, user-centered design*

General Terms: Design, Economics, Human Factors, Measurement, Performance

Additional Key Words and Phrases: Preferences, user evaluation, usability, aesthetics, auction, market value, screen design

1. INTRODUCTION

Managers, designers, and marketers would often like to know the value of various design features for the user. However, it is usually difficult to assess the value of a feature, and there may be multiple ways of doing so. Moreover,

Authors' addresses: T. Ben-Bassat and J. Meyer, Department of Industrial Engineering and Management, Ben Gurion University of the Negev, Israel; email: {sem,joachim}@bgu.ac.il; N. Tractinsky, Department of Information Systems Engineering, Ben Gurion University of the Negev, Israel; email: noamt@bgu.ac.il.

different methods may lead to different conclusions. Broadly speaking, the two most common methods used by usability researchers and practitioners to evaluate and to compare different systems and design features are objective performance measures and subjective user evaluations. Ideally, the two methods should yield similar results. But this is seldom the case [Karat 1997]. Studies have shown that measures obtained by the two methods are only moderately correlated [Nielsen and Levy 1994] if not completely disparate [Frøkjær et al. 2000]. Thus, researchers and practitioners alike are often at a loss for trying to reconcile incompatible subjective and objective measures [Meyer and Seagull 1996]. Moreover, a closer examination of the two methods' incommensurability reveals yet another potential complication. User preferences, as measured by evaluations of various aspects of the system, by expressing attitudes towards the system, or by rank-ordering different systems may not necessarily correspond to actual decisions to use or to buy one system over the other. As Einhorn and Hogarth [1981] reminded us, judgment (or evaluation) does not necessarily equal choice. Choice entails commitment and consequences that are absent from the mere evaluation of alternatives.

The assessment of a system's value for the user is further complicated by the fact that users' evaluations of systems may be based on various properties of the system besides its usability. Dimensions of the interaction that are not necessarily task oriented (e.g., content quality, fun, and arousal) can enter the users' considerations and the evaluation process [Karat 2003]. For example, in laboratory studies of users' preferences of Media player skins, Tractinsky and Lavie [2002] and Tractinsky and Zmiri [2006] found that users' choices of media player skins were evaluated on at least three different dimensions (usability, aesthetics, and symbolic value). In these studies, users' chosen skins differed from their expressed evaluations and preferences.

The weak association between performance measures and evaluation measures has prompted researchers to suggest the use of composite usability measures. Such measures estimate the overall usability of the system by taking into account both objective and subjective measures (e.g., Su [1998]; Frøkjær et al. [2000]). The composite measure should better reflect the system's value for the user and/or for the organization. However, even combining evaluations and performance measures may not suffice to portray an accurate picture of the system's real value, due to the lack of commitment and accountability inherent in the evaluation elicitation procedures. This limitation of common subjective usability measures can be alleviated if these measures are replaced or augmented by procedures that have real consequences for the user. Alternatively, it can be discounted if we can demonstrate that evaluations are good predictors of choice in the HCI context.

The main objective of this study is to demonstrate the use of an economic mechanism—the auction—to assess the subjective value of design attributes. We manipulated design features that are directly related to the performance with the system, as well as features related to the system's external appearance. The manipulations were based on the recent tendency in the HCI literature to consider hedonic and affective dimensions of the interaction, in addition to traditional HCI dimensions (e.g., Hassenzahl [2003, 2004]; Norman [2004]).

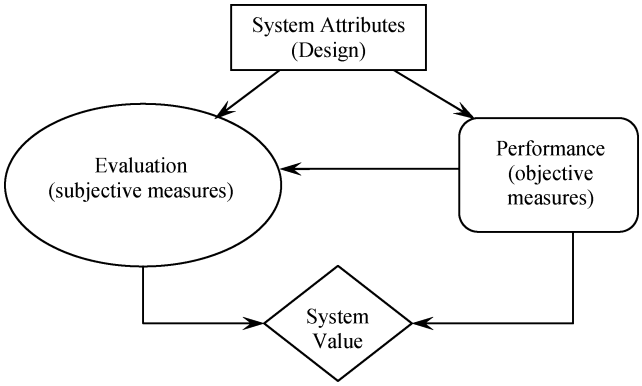

Fig. 1. Context of the study.

The study aims to determine the degree to which various design features affect performance measures, subjective evaluations as expressed in responses to questionnaire items, and the value assigned to alternatives as expressed in bids in auctions.

A major motivation for our study was the recent finding that at least in certain contexts, perceptions of system aesthetics affect users' evaluations of the system's usability [Tractinsky et al. 2000]. Aesthetics was also found to be an important determinant of user preferences of web pages [Schenkman and Jonsson 2000; van der Heijden 2003]. It is not clear, however, whether the role of aesthetics in shaping users' evaluations is not driven to a large extent by the fact that users' were not accountable for their preferences in any way. One can argue that in a laboratory environment, people may strongly weigh the aesthetic aspect of the system, because such a preference is inconsequential. In many real world situations, on the other hand, preferring a more aesthetic system over a more usable one may cause users to become less efficient, and this may have consequences in terms of job performance, promotion, earnings, and so on.

Figure 1 depicts the context of this study. It suggests that objective design attributes (e.g., display design, possible navigation paths, system messages, the set of possible interaction methods and operators) affect both actual performance (which can be measured objectively) and users' perceptions and evaluation of the system and their attitudes towards it (which can only be measured subjectively). Users' subjective evaluations of systems are based on their interaction with the system, but knowledge about their actual performance may also aid in forming those evaluations. Evaluations of two or more systems entail implicit or explicit expressions of preferences for each system relative to the others. Finally, evaluations or preferences, as well as information on performance, determine the system's value for the user. This, in turn, influences the decision of whether to use or to abandon a certain system, or whether to buy it, and if so, how much one is willing to pay for it. As mentioned above, the literature suggests that there are discrepancies between subjective valuations of systems and users' tendency to actually use the system. It is possible, though,

that the value assigned to a system reflects performance better if we use economic measures, such as bids in auctions, as a value elicitation method. This is because bids in an auction go beyond evaluations. They entail the assignment of a specific value to an object, as well as a commitment by the evaluator to pay a price, based on his or her evaluation.

In this study, we manipulate system attributes that pertain to the usability and the aesthetics of the system, and assess the effects of these attributes on subjective evaluations and on the value of the system, as expressed through users' bids in auctions. In the following section we will first discuss the issue of auctions as a measure of subjective value. We will then briefly review the literature on aesthetics and usability and their relation to user preferences. We conclude this section by discussing two contextual variables that may also affect how users value a system: monetary incentives and experience with the system. In Section 3 we present the general approach to this study and the rationale for the experimental design. Section 4 details the study's method. The results are presented in Section 5 and discussed in Section 6.

2. THEORETICAL BACKGROUND

In this section we present the major building blocks of this study: the value elicitation method (questionnaires and auctions), system attributes to be manipulated and perceived (usability and aesthetics), and contextual factors that may affect the valuation process (experience with the system and monetary incentives).

2.1 Auction Bids as Measures of Value

Questionnaires are by far the most common method for measuring user preferences. In them, users typically evaluate different systems or state explicitly which system they prefer. However, preference statements—and certainly preferences that are calculated from evaluations—do not necessarily translate to actual choice [Einhorn and Hogarth 1981]. In order to overcome this problem, it is necessary to examine additional techniques for measuring the value of systems to users. Such techniques should ideally cause users to choose among systems or to assign them some tangible price tag, based on the value of the system for the user.

In the context of evaluating system features, the assigning of values to systems should be preferred over preference rankings because valuation also provides some information about the degree to which the systems differ. This allows us, for instance, to gain some insights into the relative importance of system attributes for users by comparing the valuations of the systems with and without each attribute.

One possible technique to measure value is to use auctions in which participants bid for different systems. Bids indicate users' perceived value of the systems. The study of auctions is a very active field of economic research. For an accessible introduction to some of the main concepts in auction theory that are relevant for the current study, see Milgrom [1989]. The types of auctions that are most relevant for our purpose, and that will most likely be applicable for

the evaluation of interactive systems, are sealed-bid auctions. Here each bidder submits a bid for the item that is auctioned, and bidders cannot observe the bids of others. Usually the highest bid wins the auction. There are two different types of sealed bid auctions: First-price auctions, in which the bidder who submitted the highest bid wins the item and pays her own bid; and second-price auctions, where the highest bidder wins, but pays the second-highest bid price. Vickrey [1961] has shown that the dominant strategy in a second-price auction is for each bidder to bid her or his value of the auctioned item. In first-price auctions matters can be different, and bids of participants may be influenced by their beliefs about the bidding of other participants: bidding may be strategic. Hence, for eliciting accurate estimates of the perceived value of an interface feature, economic theory would recommend the use of second-price auctions. However, the notion of a second price auction is not intuitive, and participants may find it easier to understand first-price auctions.

Numerous experiments studied behavior in auctions in experimental settings (see Kagel [1995] for a review). Bids in experimental auctions tend to approach the actual value of the auctioned item for the bidder if auction trials are repeated a number of times and if the bidding has monetary consequences (e.g., Coppinger et al. [1980]). Therefore, auctions can be used to determine the value of goods. For instance, Fox et al. [2002] used an auction mechanism to assess the value of meat irradiation in the eyes of consumers. In such valuation studies second-price sealed bid auctions were shown to provide stable estimates of values if auctions were repeated at least three times. The method is particularly effective if the auctioned good is a market good and not an intangible entity such as health [Shogren et al. 1994].

We suggest here that auctions can be used to measure the value of interactive systems that differ in certain design features, by allowing users to bid on various versions of the system that differ in those features. The price that a system with a given feature obtains in the auction can be viewed as its market value, and the price difference between systems with and without a feature can serve as an estimate for the value of that feature. The specific bids individuals make in an auction can serve as a measure of the value of a design feature for each person. As mentioned above, such a measure may serve as a better indicator for the value of a system in situations in which system use has consequences for the user.

2.2 Design Attributes: Aesthetics and Usability

In order to compare the perceived value of systems that differ in certain features, we need to create systems that differ in salient design attributes. The different methods will be used to determine the values of each attribute. The systems should ideally also be identical in all respects, except for the target attributes. The two design attributes that we chose to manipulate in this study are the system's usability and its aesthetics. An ongoing discussion in the literature on HCI deals with the degree to which these two properties are related and whether they affect each other's evaluation. Our study may contribute to this discussion by showing whether different evaluation methods present the

same or different pictures about the relations among usability, aesthetics, and system evaluations.

The recent interest in aesthetics in HCI generates from the fact that one possible reason for the preference-performance discrepancy may be that there are other design attributes that can affect preferences besides performance [Karat 2003; Norman 2004]. One such attribute is the application's aesthetics. Research on human-computer interaction has now begun to consider the role of aesthetics in interaction design, its effects on the users, its relations with users' perceptions of other system attributes, and its impact on the overall experience of interacting with the system. This line of research suggests that aesthetics is a strong determinant of the pleasure the user experiences during the interaction. Aesthetics was found to be correlated with perceptions of the system's quality [Hassenzahl 2004], and with users' satisfaction [Lindgaard and Dudek 2003]. Schenkman and Jonsson [2000] found that beauty was a primary predictor of overall impression and preferences of web sites.

For a long time, the HCI community had considered the design dimensions of beauty and usability as independent of each other, or even as inversely related in the sense that attempts to beautify a system can reduce its usability. Recent research, however, suggests that, at least in the users' minds, these system attributes might be positively associated. Several studies have found such associations both before and after the interaction [Tractinsky 1997; Tractinsky et al. 2000; Lavie and Tractinsky 2004]. Similarly, van der Heijden [2003] found that the visual attractiveness of a web site affected users' enjoyment as well as their perceptions of ease of use and, to a lesser extent, usefulness.

These findings raise the question of how users weigh different properties when they provide an overall evaluation of a system. These cited studies imply that aesthetics has a major impact on users' preferences of interactive systems. A critical view of this proposition suggests that the importance of aesthetic considerations should depend on contextual factors such as the specific system, the user's goals and motivation, and the incentives involved. It is also possible that the relative importance of usability and aesthetics for a user may depend on the method that is used to assess user preferences. For example, because of the commitment involved in the auction process, questionnaires and economic measures may lead to different evaluations of a system, at least when the context calls for more accountability on the part of the evaluator.

2.3 Contextual Factors

In this study we include two contextual factors that may affect how users evaluate the relative importance of properties of an interactive system: The incentive structure and the experience with the system.

1. **The incentive structure.** If a user receives tangible rewards for performance, usability is likely to be more important than when the interaction with a system does not lead to rewards. The distinction between productivity-oriented (utilitarian) and pleasure-oriented (hedonic) information systems is relevant here (e.g., Hassenzahl [2003]). Utilitarian systems provide instrumental value for the user, while interaction with hedonic systems is an end

Table I. Combinations of Evaluation Method and Contextual Factors (Incentive Structure and Experience). The Five Experimental Conditions Used in the Experiment are Listed in the Appropriate Cells

Evaluation Method: Auction	Context Factor: Monetary Incentives	Context Factor: Experience with the System	
		No	*Yes*
No	*No*	Condition 1	Condition 2
	Yes	—	Condition 3
Yes	*No*	—	—
	Yes	—	Conditions 4/5

in itself. For utilitarian systems, users should value usability more than for hedonic systems. The specific task employed here places a premium on the usability aspects of the system by rewarding users for performance. Therefore, we expect that evaluations, and especially choices, will reflect greater importance of the usability dimension.

2. **The user's experience with the system.** Common wisdom, as well as usability evaluation practices, suggest that users evaluate systems differently before and after they had an opportunity to actually use them. According to this view, users become more aware of a system's usability during and after actual use. Consequently, their evaluations of the system will better reflect the system's usability after, rather than before, using it (e.g., Hassenzahl [2004]; but see also Tractinsky et al. [2000] for different findings).

3. RESEARCH APPROACH

Based on the previous section, we can vary three elements of the evaluation process: the evaluation method (whether an auction, in addition to a questionnaire, is used to evaluate the system), the incentive structure (whether users receive monetary rewards based on their performance with a system), and the user's experience with the system before providing the evaluation (whether users gain experience with the system before they evaluate it). Table I presents the eight combinations that arise from these three factors. We have used only half of the resulting combinations because four of the eight cells are of no practical or theoretical importance (e.g., using the auction method without monetary incentives or without experiencing the system). In addition, the condition that includes an auction was divided into two subconditions according to the auction type (first- and second-price auction).

Another way to view the five conditions is by considering them as being generated by cumulatively adding components to the interaction with the system and to the evaluation process. Table II presents the five different conditions and the components in each of them. In the first condition, evaluations are only based on the impressions formed by looking at the interfaces and by reading the usage instructions. At the second level, users use the system for some time and experience how it works before evaluating it. The third condition involved a monetary incentive: users' evaluations of a system take into account the fact

Table II. Components Used in Each of the Experimental Conditions

	Experimental Condition			
Evaluation Component Used	1	2	3	4/5
Visual impression and instructions	√	√	√	√
Experience with the system		√	√	√
Monetary incentive			√	√
Competitive environment (auction)				√

that their performance with that system leads to monetary rewards. Thus, in conditions 2 and 3 issues of usability increasingly become more salient.

The fourth and fifth experimental conditions employed auction mechanisms to elicit users' evaluations of the system, in addition to experiencing the system and to receiving monetary rewards. In these conditions users were asked to bid on systems that differ in usability and aesthetics and are awarded one of the systems, based on their bids. They were rewarded at the end of the experiment according to their performance, after having to pay their bid for the system they bought in the auction. Thus, users had to decide on the value of each system and experienced the consequences of their valuations through the price they paid for the system and the benefits from using it. In the auction conditions, the users' bids provide us with measures of their relative preference of the different systems, which can be compared to the questionnaire-based evaluations. Also, the prices at which different systems are sold in these auctions provide some estimate of the market value of the different features, in addition to the bids as measures of individual valuations.

3.1 Research Questions

Based on these factors and their potential role in users' evaluations of interactive systems, our study addresses five main questions:

1. *Aesthetics vs. Usability*. Previous studies have shown that subjective evaluations of usability and aesthetics are correlated. The sources for these correlations are yet unclear. Tractinsky et al. [2000] suggested that the system's beauty affects perceptions of its usability: "beautiful is usable." Hassenzahl [2004], on the other hand, suggests that such a relationship exists only when it is impossible to find in the system pool, systems that are both beautiful and not usable, and systems that are both usable and ugly. In our study, we independently manipulate both the aesthetics and the usability of the system in an attempt to approach the conditions under which Hassenzahl sees the boundaries of the "beautiful is usable" phenomenon. The independent manipulation will facilitate testing the degree to which each of these two aspects of the system (aesthetics and usability) affect perceptions of the other aspect.
2. *Effects of Experience on Evaluations*. The experimental conditions differ in the amount of experience users have with the system. Hence we can evaluate how actually using a system affects its evaluation. While Tractinsky et al. [2000] found no effect of experience with a system on subjective evaluations,

Hassenzahl's [2004] results indicate that using the system helped untangle the associations between usability and beauty. However, neither of these studies (nor any other study that we are aware of) manipulated usability and aesthetics independently. Since we did so in the current study, we expect a clearer answer to the question of whether experiencing a system mitigates the initial perceived association between usability and aesthetics.

3. *Effects of Rewards on Evaluations*. Previous studies on the relative importance of usability and aesthetics tended not to reward users for their performance. The use of actual monetary rewards may change perceptions of usability and aesthetics (as well as the relation between the two aspects).
4. *Evaluation vs. Bidding*. In this study, evaluations of system usability and aesthetics will be collected by a questionnaire and by users' bids in auctions. Thus far, research in the field of HCI mainly used questionnaires to assess user evaluations of systems. The auction mechanism is different in that it forces the user to form an explicit evaluation of each system and to bid accordingly. The literature suggests that evaluations and willingness-to-pay do not always coincide. Hence the different methods for measuring user preferences may provide different evaluations of the systems.
5. *Effects of Auction Method*. In this study we will use two different types of auctions (first-price auction vs. second-price auction). We are interested in learning whether the type of auction affects users' preferences, choices, and the amount of money they bid on each system.

4. METHOD

4.1 Participants

Participants in the study were 150 engineering undergraduate students. These students were not exposed to aesthetics considerations in artifact and interface design during their academic studies. The participants were assigned in equal numbers to the five experimental conditions. Due to technical malfunctions and after screening of outliers (observations exceeding 2 standard deviations over or under the mean result were discarded), 143 of the 150 observations were included in the analysis.

4.2 Materials

The experimental program was written in MS Visual Basic. The system simulated a computerized phone book, which included a screen and a virtual keyboard. Groups of 15 participants were seated in a computer classroom. All computers had 17″ screens that were set to a resolution of 600x800 pixels. Participants interacted with the system via a mouse. Keyboard input was disabled in order to limit the possible effects of typing skills.

4.3 Experimental Design and Manipulations

The experimental conditions were based on an incomplete design of three between-subjects factors (evaluation method, monetary incentives and

Table III. The Five Interactive Systems Used in the Experiment. Designs Differ in Usability (Low, Medium, High) and Aesthetics (Low, High). The Low Usability Design Was Only Presented with Low Aesthetics, Serving as a Baseline System

		Aesthetics	
		Low	High
Usability	Low	Baseline	
	Medium	1	2
	High	3	4

experience) and two within-subjects factors (usability and aesthetics). The three between-subjects factors created the five experimental conditions described in Tables I and II. In all experimental conditions, participants received instructions on how to use the different systems. In four of the five conditions, participants used the systems (see Table II). Three conditions involved monetary rewards, based on the users' performance. Finally, in two conditions, users participated in auctions in which they bid on different systems. This experimental design is cumulative: in each condition components of the lower order conditions were maintained, and an additional component was added. Condition 1 included only one component, whereas conditions 4 and 5 included all 4 components. The difference between conditions 4 and 5 was in the type of auction used: First-Price auction in condition 4 and Second-Price auction in condition 5. In all experimental conditions, the participants indicated their view of each system's usability and aesthetics using multiple-item scales. Detailed descriptions of all conditions are presented in Subsection 4.6.

The two within-subjects factors generated four systems that differed in usability (medium and high) and aesthetics (low and high). A fifth, baseline system had low aesthetics and low usability (see Table III for a description of the five systems). The baseline system was used during the training stage of the experiment and was intended to enhance the competition in the auction conditions. Each of the four experimental systems was identified by logo and a name of a flower—cyclamen, rose, anemone and daffodil—in order to help the participants distinguish between the systems.

It is not easy to manipulate the aesthetic and the usability factors independently. Not only did studies find users' perceptions of these attributes to be correlated (e.g., Lavie and Tractinsky [2004]), but it is obvious that changes to the interface's design that are targeted at changing its aesthetics may also affect the usability of the system, and vice versa. To manipulate usability, we created identically looking systems that only differed in the number of keystrokes required to perform a task. This resulted in longer task completion times and slightly higher probability of errors for systems that require more keystrokes. It can be argued that by that we have manipulated only certain usability aspects, whereas usability is a multifaceted concept (e.g., Shackel [1991]; Karat [1997]). However, our experimental systems manipulated some of those facets

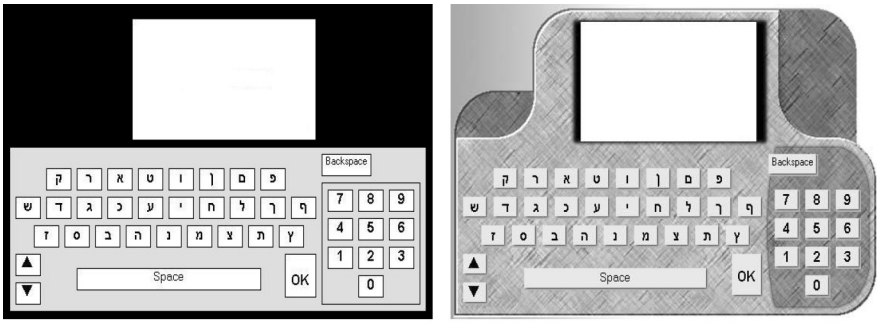

Fig. 2. The low and high aesthetics designs that were used in the experiment.

(e.g., ease of use, completion time, error prevention) while maintaining the other aspects (e.g., learnability) basically intact. Thus, the overall usability level was affected, even though certain usability aspects remained the same across all usability levels.

For the manipulation of aesthetics, we created a number of versions of the same system that were identical in terms of functionality, in terms of size and clarity of the interface's objects, and in terms of the layout of those objects on the display. The versions did differ, however, in their graphical design. Thus, in order to prevent cross effects of aesthetics and usability, only minimal (mainly decorative) aesthetic changes were made. These changes were mainly stylistic in nature and were mostly related to the background of the displays, rather than to the interactive components. (It is worth noting that the constraints on the aesthetic manipulation of the systems resulted in a fairly limited manipulation of this factor.) We then pilot-tested the aesthetics of the different systems and conducted the main study using the two systems that were rated the farthest apart.

4.4 Pilot Testing

Both usability and aesthetics were pretested in pilot studies:

The Aesthetic Factor. In a pilot study, six designs of the same system were presented to 31 participants. All designs were based on gray shades, in order to avoid uncontrolled color influences. A four-item scale (with end points at 1 and 7) was used to assess the perceived aesthetics of each design. Out of all six systems, two were selected (see Figure 2)—one as a low aesthetics system (mean aesthetics rating of 3.17) and the other as a high aesthetics system (mean aesthetics = 3.75). Because of the constraints on the aesthetic manipulations (see Section 4.3), the difference between the aesthetics ratings for the two systems was small. Still, it was highly significant $\{t_{(30)} = 4.35, p < .001\}$.

The Usability Factor. Usability was manipulated through the number of steps required to perform the experimental task of entering data into a phone book. The phone book simulator included a menu system. In the high usability system, the task did not require going through menus, and users could enter the

data directly. The medium usability system contained one menu, and the low usability system contained two menus, and its numeric keypad was arranged horizontally at the top of the keyboard rather than as a group at the right side of the keyboard. A GOMS analysis of the three systems predicted that the high usability system would be faster by about 24% than the medium usability system, which, in turn, would be faster by about 37% than the low-usability system.

In a pilot study, 12 participants were asked to perform the experimental task for 6 minutes on each of the three systems (one for each usability level) in a random order. There were significant differences between all three systems. Performance was consistent with the predicted usability level of the system, with means of 12.25, 17.75 and 23.00 items entered correctly in 6 minutes for the low, intermediate and high usability systems, respectively $\{F_{(2,12)} = 204.6, p < .0001\}$. Results of the pilot study also showed some learning over time, but performance levels became stable after about 4 minutes. We therefore decided to test performance in subsequent parts of the study for periods of 5 minutes.

4.5 Task

The experimental task was to enter data into a simulated computer based phone book. Each data item consisted of a name (in Hebrew), made of 4–6 letters and a number, made of 7 digits. The program used a name database containing 210 items, which included names made of 4, 5, and 6 letters in equal proportions. The items were presented in random order. A new item appeared when the user clicked the "OK" button after entering the name and number for the previous item. The goal was to enter as many data items as possible in a limited period of time. Users were paid based on the number of items they correctly entered into the system. An item was only considered correct if the user typed both the name and the number correctly. No indication was given if an item was typed incorrectly.

4.6 Procedure

The experimental session included 4, 8, or 10 stages, depending on the experimental condition (see Table IV). In the first stage of all conditions, the program displayed a demographic questionnaire. At the end of the experiment all participants responded to a questionnaire that measured perceived usability and aesthetics of each system. The questionnaire was presented for each of the four experimental systems in the order in which the user had encountered the systems during the experiment.

Condition 1—*Instructions Only*. Participants were asked to read instructions describing the use of each of the four experimental systems twice. In addition, they were shown the data entry task screen of each system, but they could not operate the systems. The order in which the instructions for the different systems were presented was individually randomized for each participant. In this condition, participants were not presented with the baseline system. After reading the instructions for the four experimental systems, participants filled out the concluding questionnaire for each of the systems.

Table IV. Procedure Used in Each of the Five Experimental Conditions

	Condition				
Stage	1	2	3	4	5
1 – Completing a demographic questionnaire	√	√	√	√	√
2 – Reading usage instructions and entering data for 3 minutes with the baseline system (learning stage)		√	√	√	√
3 – Entering data for 5 minutes with the baseline system (performance stage)		√	√	√	√
4 – Reading instructions on how to use the four experimental systems	√	√	√	√	√
5 – Entering data for 2 minutes with each of the four experimental systems (learning stage)		√	√	√	√
6 – Reading instructions on using the four experimental systems (second time)	√	√	√	√	√
7 – Entering data for 5 minutes with each of the four experimental systems (performance stage)		√	√	√	√
8 – Participating in 5 successive auctions in order to buy one of the four experimental systems				√	√
9 – Entering data for 5 minutes with the system they won in stage 8				√	√
10 – Completing a questionnaire on perceived usability and aesthetics for each system	√	√	√	√	√
Total experience with each experimental system (minutes)	0	7	7	12	12
Monetary incentive	No	No	Yes	Yes	Yes
Competitive environment (auction)	No	No	No	Yes	Yes

Condition 2—*Experience*. Following the instructions, the experimental program displayed the data entry screen. During the first part of the experiment, the participants performed the experimental task with the baseline system for a 3-minute learning period, followed by a 5-minute performance period. Then participants used the four experimental systems. There was a learning period of 2 minutes with each of the four systems (presented in individually randomized order). After participants had completed the learning periods with all systems, they moved on to performance periods in which they used each system for 5 minutes. Systems were presented in the performance period in the same order as in the learning period. After performing the task with all four systems, participants filled out the concluding questionnaire for each system in the same order they had encountered them before.

Condition 3—*Monetary Incentives*. The experimental procedure in this condition was similar to the Experience condition (Condition 2). However, after completing the performance stage with each of the four experimental systems (stage 7) participants were notified of their gains, based on their performance with all 5 systems. They received 0.5 NIS (about US $0.10) for each correct item (name + number) they had entered. The participants were guaranteed a minimum payment of 30 NIS (about US $6.50).

Condition 4/5—*Competitive Environment*. The first 7 stages of this condition were identical to those of Condition 3 (Monetary Incentives). At the end of the performance stage, the program presented the cumulative gain from all 5 systems. However, the payment for performance was lower in these conditions, only 0.1 NIS (about US $0.02) for each correct item. In the next stage of the

experiment, the users participated in 5 successive auctions in which they bid on the four experimental systems. The multiple auctions manipulation was based on previous findings (e.g., Coppinger et al. [1980]; Roth and Ockenfels [2000]) that showed that in multiple successive auctions the bidders raise their bids and approach the actual value of an item. The auctions took place in groups of five participants. The participants did not know who their competitors were, since 3 groups were present in the laboratory simultaneously. Out of the 5 auctions, one was randomly chosen to determine the allocation of systems to participants in the subsequent performance stage.

Auction results were determined as follows: First, the highest bid in the auction was found, and the bidder received the system for which she bid. Then the highest bid for the remaining systems was determined among the bids of the bidders who still had not won a system. The person with the highest bid received her system of choice. This process continued for all four experimental systems. The fifth participant, who had not won any of the systems, received the default (i.e., baseline) system without having to pay for it. If two participants offered the same bid for a system, one of them was randomly chosen as the winner.

Conditions 4 and 5 differed in the type of auction used. Condition 4 was a First-Price auction, where the bidder who offered the highest price wins and pays the sum she offered. Condition 5, on the other hand, was a Second-Price auction, where the bidder who offered the highest price wins, but pays the second highest price offered in the auction. In both auction types, the bids were sealed and each user was asked to offer prices for all four experimental systems. The auction screen in the experimental program presented the four systems on which participants could bid, and next to each system was a field in which participants could enter their bids for each of the systems. Participants could return to the instructions page if they felt they needed additional information. After entering the values for all systems, participants submitted their bids and moved on to the next stage of the experiment. Bids could be any value between 0 and 50 NIS (about US $11). At the end of each auction, participants were informed which system they had won in the auction, how much they would have to pay for it, and were reminded of their bids. In the Second-Price auction condition, participants were also notified of the sum they would have to pay for the system they won (the second highest bid).

In the next stage, one of the five auctions was randomly chosen. Based on that auction's results, systems were assigned to participants who used them to enter data for another 5 minutes. This time, the payment for performance was 3 NIS (about $0.70 US) for each correct item. The participants' financial compensation at the end of the experiment was the sum of their rewards for correctly entered data during the whole experiment, minus their payment for the system (as determined through the auction).

4.7 Dependent Measures

The variables of interest in this study were objective performance measures, subjective preference measures, and auction bids for each of the four

experimental systems. Each of these variables was tested as a function of usability and aesthetics (two within-subjects factors) in each of the five different experimental conditions (a between-subjects factor).

At the end of each experimental condition, the participants filled out a Likert-type questionnaire in which they indicated their agreement with statements regarding the aesthetics and the usability of each of the four experimental systems. The scales' end points were 1 ("strongly agree") and 7 ("strongly disagree"). While there are comprehensive scales for the assessment of usability, these were too long for our purpose; the study was already long enough, and such a questionnaire would have consumed too much time and energy from the participants. Also, our experimental systems were relatively simple and a detailed usability assessment was not required. Instead, we employed a general scale of perceived usability that mainly captures the usability aspects manipulated in this study. Such a scale reflects the convenient shortened form for the definition of the concept: 'the capability to be used by humans easily and effectively' [Shackel 1991, p. 26]. The items used in the usability scale were adapted to this study, based on the validated usability scale used by Lavie and Tractinsky [2004] in the context of Web pages. The items were: "using the system was comfortable"; "the system was easy to use"; "the system functioned well"; and "interactions with the system were quick." For the same reasons, we adopted a short and general aesthetics scale. The three items on aesthetics were: "the system is beautiful"; "the system is aesthetically designed"; and "I like the design of the system." This scale was previously used and validated in a series of unpublished works by the third author.

A Cronbach Alpha reliability score was computed for the evaluations of each system across all experimental conditions. Thus, for each perceived measure (aesthetics and usability) we had four reliability scores. The reliabilities of the four-item usability scale used to evaluate the four systems ranged from .83 to .91; the reliabilities of the three-item aesthetic scale ranged from .90 to .93.

5. RESULTS

5.1 Manipulation Check

The results presented in this subsection refer to the performance stage of the experiment in Conditions 2–5 (Stage 7 in Table IV). Performance was measured by the number of items entered into the system in the allotted time (5 minutes). The baseline system was used as part of the experimental procedure only in order to create a default system in the auction conditions (conditions 4, 5). Thus, it was not included in the main analyses, but it was still important to ensure that it led to the worst performance.

We first analyzed the performance levels in all five systems (including the baseline system), across 4 experimental conditions (Conditions 2–5, because Condition 1 did not involve actual performance). This was a two-way analysis of variance (ANOVA) with System Type as one within-subjects factor (with five categories corresponding to the five systems) and Condition as a between-groups factor. Results showed a significant effect of the system

Table V. Performance Means (and Standard Deviations) as Measured by the Number of Items Entered in 5 Minutes Using Five Systems, Across Four Experimental Conditions

		Aesthetics	
		Low	High
Usability	Low	13.70 (2.04)	NA
	Medium	21.07 (2.98)	20.50 (2.95)
	High	26.08 (3.74)	25.51 (3.65)

$\{F_{(4,452)} = 1399.403; p < .001\}$, no effect of the condition factor $\{F_{(3,113)} = 2.114; p = .102\}$, and no interaction effect $\{F_{(12,452)} = 1.625; p = .074\}$. In addition, pairwise comparisons showed a significant difference between the baseline system and the four experimental systems (see Table V).

Performance was then examined for the four experimental systems (without the baseline system), based on the manipulation of the two independent variables: usability (medium and high) and aesthetics (low and high). These variables served as within-subject factors and the experimental condition (four levels) served as a between-subjects factor. The three-way ANOVA revealed a strong effect of usability $\{F_{(1,113)} = 1381.774; p < .001\}$, indicating that the usability manipulation succeeded. There was also a small, but significant effect of aesthetics $\{F_{(1,113)} = 23.35; p < .001\}$. Performance with low aesthetic systems was slightly better than with high aesthetic systems. No other main effects or interactions were significant.

Thus, in reference to Figure 1, the results indicate that the manipulation of the systems' design attributes yielded the expected objective (i.e., performance) results. We now turn to analyzing the subjective (evaluation) measures.

5.2 Effects of Experimental Factors on the Subjective Measures

Recall that subjective measures were recorded by questionnaires for each system at the end of each of the 5 experimental conditions. This subsection analyzes the effects of the experimental factors (usability and aesthetics as within-subjects factors and condition as a between-subjects factor). The analyses conducted on the data are based on manipulations of three independent variables. System usability (two levels: medium and high) and system aesthetics (two levels: low and high) were within-subjects factors. Experimental condition (five levels: instructions only, experience with the system, monetary incentives, and two types of auctions) was a between-groups factor.

Perceived Usability. A three-way analysis of variance was conducted in order to examine, across all participants, the influence of usability, aesthetics and experimental condition on perceived usability. There was no significant main effect of the experimental condition on perceived usability, nor were there

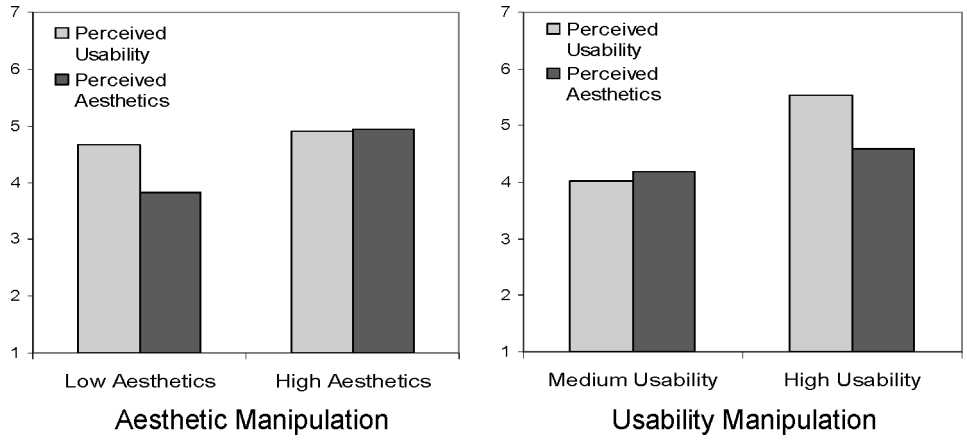

Fig. 3. Questionnaire-based evaluation of the system (i.e., perceived usability and perceived aesthetics) as a function of experimental manipulations of the systems' attributes: aesthetics (left panel) and usability (right panel). Scale end points: 1 (low) to 7 (high).

interaction effects between experimental condition and system usability, or system aesthetics on perceived usability. Thus, experience with the system or the bidding on systems in auctions did not change the perceived usability of the systems.

Figure 3 shows the effect of manipulating the system's design attributes on subjective evaluations of usability and aesthetics. The right panel in Figure 3 depicts the effects of the usability manipulation; the left panel depicts the effects of the aesthetics manipulation. The main effect of usability on usability ratings was significant $\{F_{(1,137)} = 162.089; p < .001\}$. Intermediate usability systems were rated as less usable than high usability systems (with means of 4.02 and 5.54, respectively). A significant effect was also found for the aesthetics factor $\{F_{(1,137)} = 8.104; p < .005\}$. More aesthetic systems were perceived to be slightly more usable, with perceived usability means of 4.66 and 4.90 for the low and high aesthetics systems, respectively. These perceptions contradict the actual performance measures, where the low aesthetics system actually brought about slightly better performance.

Perceived Aesthetics. A three-way analysis of variance was used to test the influence of usability, aesthetics and condition on perceived aesthetics. The effect of the experimental condition was not significant. The analysis revealed significant main effects of aesthetics on aesthetics ratings $\{F_{(1,137)} = 45.957; p < .001\}$ and on usability ratings $\{F_{(1,137)} = 31.937; p < .001\}$. As can be seen in Figure 3, participants tended to rate the high aesthetics system as more aesthetic than the low aesthetics systems (means of 4.94 and 3.83, respectively). They also rated the high usability system as more aesthetic than the intermediate usability systems (means of 4.59 and 4.18, respectively).

There was a significant interaction of the experimental condition with the usability factor $\{F_{(4,137)} = 3.518; p < .009\}$, but this interaction is difficult to interpret. Overall aesthetics was always judged as higher for high usability

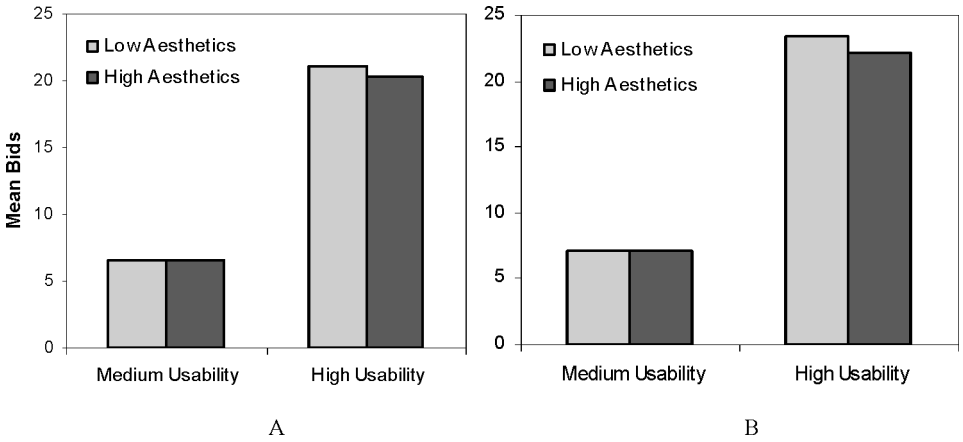

Fig. 4. Mean bids across both types of auctions for five repeated auctions (a) and mean bids for the last auction (b), for systems differing in aesthetics and usability.

systems, except for the condition with monetary incentives (Condition 3). In this condition the aesthetics ratings for high, and intermediate usability systems were very similar.

Overall, neither experience nor monetary incentives had a substantial effect on users' evaluations of the systems' usability and aesthetics. In all experimental conditions, the two properties of a system were perceived similarly.

5.3 Auction Bids

In the auction conditions (First and Second-Price auctions), the participants were asked to offer bids for each of the four experimental systems, in five successive auctions. To control for individual differences, the bids were standardized for each participant according to his/her mean bid across all systems. However, the analyses of the standardized bids yielded almost identical findings to the analyses of the raw bids. Therefore, we chose to use the raw data for the following analyses.

Roth and Ockenfels [2000] demonstrated that in repeated auctions, bidders increase their bids, and they approach over time the actual value of the auctioned item for the bidder. Therefore, a separate analysis will be presented for the mean bids across all five auctions and for the bids in the fifth (last) auction.

Relative Importance of Aesthetics and Usability in Different Auctions. Mean bids across five auctions. The mean bids in First and Second Price auctions are presented in the left panel of Figure 4 for the four experimental systems. A three-way analysis of variance with repeated measures was conducted with auction type (first vs. second price) as a between-subjects factor, and Usability and Aesthetics as within-subjects factors. The results revealed a significant effect of usability in the expected direction $\{F_{(1,56)} = 1971.153; p < .001\}$ and no significant effect of aesthetics $\{F_{(1,56)} = 0.035; p = .853\}$. There was no Aesthetics X Usability interaction effect $\{F_{(1,56)} = 1.752; p = .191\}$ and no interactions with the auction type factor.

Bids in the last (fifth) auction. Similar to the main effects found in the analysis of all five auctions, the usability factor had a significant effect on the last auction $\{F_{(1,56)} = 319.122; p < .001\}$, indicating that usability affected the bids in the expected direction (see the right panel in Figure 4). The aesthetics factor had no significant effect $\{F_{(1,56)} = 0.707; p = .404\}$, nor did any of the interactions.

A comparison between average bids and last (fifth) auction bids showed that, congruent with Roth and Ockenfels's [2000] findings, bids in the last auction were higher than bids across all auctions $\{F_{(1,56)} = 15.115; p < .001$; the relevant means are shown in Figure 4$\}$. The difference was larger for more usable systems. Neither the main effect of auction type (first- or second-price) nor any of its interactions, were significant. Hence it seems that participants responded to both types of auctions similarly.

Market Value Analysis. The specific bids that individuals offer in an auction measure the value of a certain system for a person. A different measure obtained with auctions is the system's market value. This value is the price that is actually paid in order to buy the system. The actual purchase prices can be used to measure the market value of usability and aesthetics in the experimental context of this study. Cases here were the 12 auction groups (six groups in each auction type). For each group we computed the market prices for the different interfaces in each of the repeated auctions.

In order to determine the effects of usability and aesthetics on the systems' market price we conducted a four-way ANOVA with the system's usability, aesthetics and auction number as within-subjects factors and the auction type as a between-subjects factor. The analysis revealed a significant effect of the system's usability on its market value $\{F_{(1,10)} = 112.48; p < .001\}$, but no effect of aesthetics $\{F_{(1,10)} = .001; p = .974\}$. The market value of more usable systems was significantly higher than the value of less usable systems (see Figure 5). In addition, a significant effect was found for the auction type (First or Second-Price auction). Results show that market prices were higher in the First-Price auction compared to the Second-Price auction $\{F_{(1,10)} = 8.03; p < .018\}$. This difference in auction type was especially pronounced in the more usable systems.

There was also a significant effect of the auction's order $\{F_{(4,40)} = 3.067; p < .027\}$. As shown in Figure 5, prices increased as auctions progressed. A significant interaction of Auction Number X Auction Type $\{F_{(4,40)} = 3.036; p < .028\}$ indicates that this was the case especially for the usable systems in the Second-Price auction.

6. DISCUSSION

We conducted an experiment on users' evaluation and valuation of systems with different levels of aesthetics and usability. The experimental conditions differed in the experience participants gained with the systems before evaluating them and the existence of monetary incentives. System evaluations were collected with questionnaires and through the bids in two types of auctions. Ideally, system evaluations should correspond to the performance with the system, and different evaluation methods should yield identical results if they tap the same

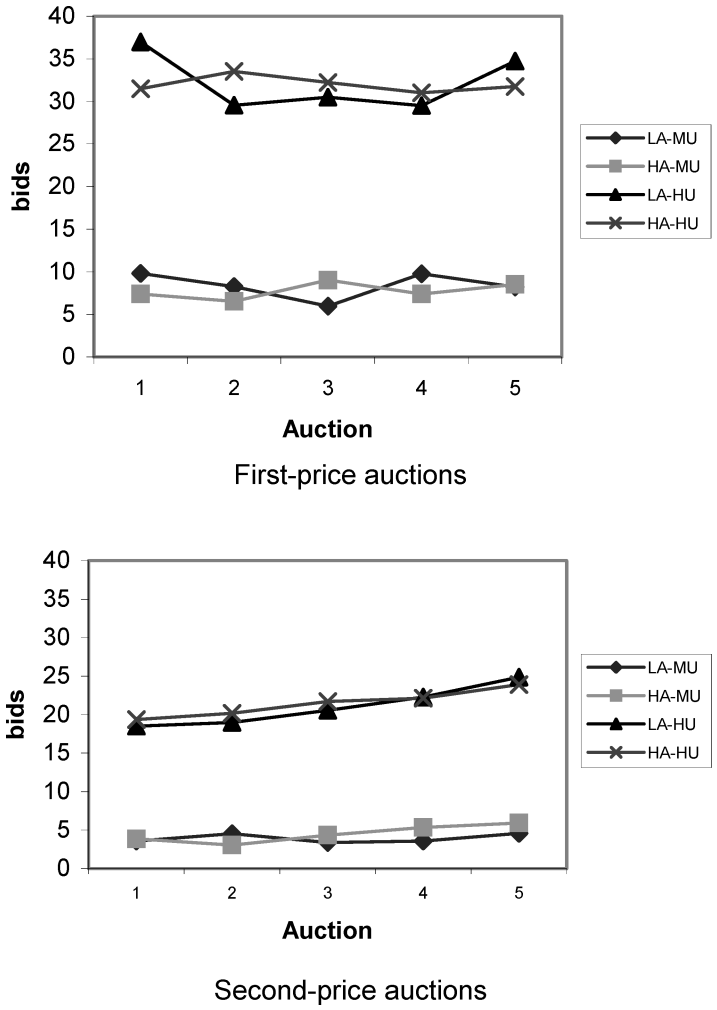

Fig. 5. The market price for 4 experimental systems in 5 successive auctions, first-price auctions (upper panel) and second-price auctions.
Legend: LA- Low Aesthetics; HA- High Aesthetics; MU- Medium Usability; HU- High Usability

basic evaluation mechanism. This was not the case. We begin this section with a discussion of results that pertain to the comparison of the different preference elicitation methods and conditions. We then discuss findings regarding the effects of the systems' attributes (usability and aesthetics) on users' perceptions of these attributes.

6.1 Comparison of Preference Elicitation Methods

Experience with the system and the presence or absence of monetary performance rewards had no effect on users' evaluations of the systems as expressed in their responses in the questionnaire. Hence, it seems that evaluations of usability and aesthetics are formed after only a brief exposure to the system.

This may demonstrate again, the importance of first impressions in shaping users' attitudes towards interactive systems [Lindgaard et al. 2006; Tractinsky et al. 2000].

Questionnaire-based evaluations differed clearly from bids for the systems in auctions. Most notably, questionnaire-based evaluations were significantly influenced by the system's aesthetics, whereas auction bids showed no such effects. Participants in this study were unwilling to pay higher prices for more aesthetic systems. Only the systems' usability affected users' bids in auctions. Consequently, market value, operationalized as the sales price in the auctions, reflected the objective usability of the systems but not their aesthetics. We therefore conclude that of the two evaluation methods employed here, auctions yielded expressed preferences that were more consistent with the nature of the experimental task, with the purpose of the experimental system (which stressed productivity and efficiency), and with actual performance.

Our results show that bidding in first- and second-price auctions is quite similar. Thus, both auction types can be used to measure user preferences. The choice of the auction method employed may depend on a method's relative simplicity and understandability to users. Second-price auctions have, of course, the advantage that the theoretically optimal bidding behavior is to bid the perceived value of the good. We conducted multiple repeated auctions based on Roth and Ockenfels' [2000] finding that users tend to raise their bids in successive auctions. Our results confirm Roth and Ockenfels' findings: bids in the last (fifth) auction were higher than the average bids across the five auctions, and mean bids increased in successive auctions. Future use of auctions to elicit the market value of systems or features should take this finding into account and use multiple auctions to assess preferences.

The results suggest a possible explanation for the weak relationship found in the literature between system performance and user preferences [Frøkjær et al. 2000; Nielsen and Levy 1994; Su 1998]. Conventional methods of preference elicitation are based on users' judgments of systems when these judgments have no consequences for the users. When this is the case, users' evaluations may reflect the influence of factors that are unrelated to actual system behavior. These factors may range from methodological issues, such as a tendency not to use the evaluation scale's extreme points, to more substantive issues, such as allowing other system attributes (e.g., aesthetics) to influence (whether consciously or not) perceptions of the system's usability.

Our results reinforce the observations of Tractinsky and Lavie [2002] and Tractinsky and Zmiri [2006] regarding the limitations of questionnaire-based evaluations to capture users' actual preferences. In these studies, which were conducted in the context of entertainment systems, users expressed preferences for more usable systems, but their actual choice implied preferences for more aesthetic systems. Our experiment stressed performance. Still, questionnaire-based evaluations implied that users were influenced by both usability and aesthetic factors. However, only the system's usability affected bids in the auctions. Together, these studies demonstrate the importance of assessing users' preferences with methods that go beyond conventional questionnaires.

Our study demonstrates the effectiveness of using nontraditional evaluation methods, but numerous methodological issues remain to be examined in future studies. For instance, we did not study bidding for systems with which users did not gain experience. Such a condition may help to clarify the effects of users' first impressions. Another potentially useful research direction is the validation of controlled experimental results with market success measures of products or product features. To what extent will the expressed value in auctions predict market behavior better than other methods of system evaluation, such as questionnaires? Also, other economic measures besides auctions should be explored.

6.2 Effects of System Attributes on Users' Perceptions

The relation between usability and aesthetics has become the focus of a recent research stream in HCI [Hassenzahl 2004; Lavie and Tractinsky 2004; Lindgaard and Dudek 2003; Tractinsky et al. 2000]. The existing studies have shown interdependence between usability and aesthetics, but the magnitude of these relations and their causality are not yet clear. In the current study we manipulated usability and aesthetics independently. This makes it possible to draw some conclusions on the direction of effects (does usability affect perceived aesthetics? Does aesthetics affect usability? Do both aspects of a system affect each other?). The manipulation also approaches (but clearly does not reach) Hassenzahl's [2004] boundary condition of a stimulus set that includes both usable and ugly systems, as well as beautiful and unusable systems. Although such a condition is practically useless, it is of theoretical interest. For obvious practical reasons we did not create such a state: firstly, designing an utterly ugly system would probably hinder usability as well, something we were determined to avoid because of the need to preserve the aesthetics-usability independence. Secondly, a completely "not usable system" would not allow users to complete their tasks, thus rendering the entire experiment meaningless.

At least two characteristics of the current study should limit the effect of aesthetics on perceived usability. First, the study clearly stressed and rewarded performance, which should have raised the importance of usability in determining users' preferences. Second, the manipulation of the aesthetics factor was not very strong. In addition to the design constraints mentioned in Section 4.3, all systems were displayed in monochrome, and the classification of systems to either high or low aesthetic levels was based on another sample's consensus, rather than on the tastes of each of the participants in this study [cf. Tractinsky 2004]. Thus it was reasonable to expect users to ignore the aesthetic aspects of the systems and to rely only on their usability.

However, the responses to the questionnaire showed that perceived usability and aesthetics were interdependent. Although performance was somewhat lower with the more aesthetic systems, users still perceived them as more usable than the less aesthetic systems. This finding is congruent with the idea that aesthetic design makes things appear more usable [Norman 2004; Tractinsky et al. 2000]. At the same time, aesthetics evaluations of the systems were higher

for the more usable systems. This relation indicates the effect of usability on perceived aesthetics. It is consistent with Hassenzahl's [2004] proposition that usable things appear more beautiful.

In contrast to the responses to the questionnaire, auction bids and the resultant market price measure demonstrated the effect of performance on users' preference. The market value of usable systems was significantly higher than the value of less usable systems, and the system's aesthetics had no effect on that value. Thus, it may well be the case that by being held accountable for their evaluations (as in this study, by having to pay for their bids), users' preferences resemble more closely the context within which the evaluations are made. In this study, the context clearly entails a priority for the usability aspects of the system. In other contexts (e.g., leisure systems, gaming) bids may reflect different priorities for various design attributes. Nonetheless, because of the intrinsic advantages of the auction-based measures, we suspect that even in other contexts such measures may reflect users' preferences at least as well as conventional measures.

6.3 Limitations

Researchers are often faced with the need to trade off considerations of external and internal validity. The emphasis in this study was on controlling the experimental conditions to increase internal validity. Consequently, certain aspects of external validity were compromised. For example, the experimental systems were relatively simple. They are not representative of many real-world applications in terms of their complexity and range of features. Still, it is possible to look at the systems used in this study as representing a small application or as a component of a decomposed larger system. But clearly, the major issue here is that the number of uncontrolled variables in large systems would not allow us to control the study's independent variables. Thus, while this study was able to demonstrate how users' evaluations are influenced by system attributes and value elicitation method (but not by monetary incentives and experience), replications of the results in more ecologically valid contexts are required. For example, in our study the system was relatively simple, and thus usage instructions may have given the participants enough information regardless of experiencing it. In more complex systems, instructions might not suffice to create the same impression on the user that actual use of the application creates.

Another limitation of the study is that only general evaluations of usability were measured, while more refined and detailed measures can be used in real-life situations. The usability scale was based on summative (as opposed to formative) evaluations, meaning that it captured only the final evaluation rather than the process by which they were formed. Thorough investigations of all aspects of the systems' usability and their effects on users' interaction with the systems are important, but this was not feasible within the framework of this study. More research is needed on the relation between aspects of usability and aesthetics and users' valuations of interactive systems as expressed through different elicitation methods.

7. CONCLUSIONS

This study aimed to compare different methods of system evaluation, to examine the effects of aesthetics and usability on user preferences, and to provide additional data on the relation between these two aspects of the systems. The two evaluation methods yielded different results. With questionnaires, the familiar incongruity between subjective preferences and objective performance was evident (as demonstrated by Nielsen and Levy [1994]. Users did not necessarily prefer the system that allowed them to reach the highest level of performance, but also considered aesthetics. This finding appeared even after users had gained experience with the system and received monetary incentives based on performance. Thus the interdependence of perceived aesthetics and usability in questionnaires seems to be a rather robust finding. However, when users evaluated systems by bidding on them in auctions, a strong positive correlation between performance and evaluations was evident, and aesthetics had no effect.

The results of this study confront usability engineers with a trade-off. The study reveals differences between traditional evaluation methods for assessing users' preferences and the auction method. On the one hand, the results of the auction method are more commensurate with performance measures and with the evaluation context. The auction method effectively solved the performance vs. preference paradox. On the other hand, using auctions to elicit preferences is clearly less convenient. The choice of the evaluation method should thus depend on the specific context in which an evaluation is needed. For example, if correctly predicting users' future behavior is of paramount importance, then it may be advisable to use auctions or similar economic measures. Organizations that are interested in accurate predictions may need to develop the infrastructure for conducting auction-based evaluations as another tool in the arsenal of usability experts. Such infrastructure may reduce the marginal costs of conducting auction-based usability testing, and can smooth the way to using this method.

REFERENCES

COPPINGER, V., SMITH, V., AND TITUS, J. 1980. Incentives and behavior in English, Dutch and second-bid auctions. *Econ. Inquiry 18*, 1–22.

EINHORN, H. J. AND HOGARTH, R. M. 1981. Behavioral decision theory: Processes of judgment and choice. *Ann. Rev. Psych. 32*, 53–88.

FOX, J. A., HAYES, D. J., AND SHOGREN, J. F. 2002. Consumer preferences for food irradiation: How favorable and unfavorable descriptions affect preferences for irradiated pork in experimental auction. *J. Risk and Uncertainty 24*, 75–95.

FRØKJÆR, E., HERTZUM, M., AND HORNBÆK, K. 2000. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*. Hague, The Netherlands, ACM Press, New York, NY (April), 345–352.

HASSENZAHL, M. 2003. The thing and I: Understanding the relationship between user and product. In *Funology: From Usability to Enjoyment*, M. A. Blythe, A. F. Monk, K. Overbeeke, and P. C. Wright, Eds. Kluwer Academic Publishers.

HASSENZAHL, M. 2004. The interplay of beauty, goodness and usability in interactive products. *Hum. Comput. Inter. 19*, 4, 319–349.

JORDAN, P. W. 1998. Human factors for pleasure in product use. *Applied Ergonomics 29*, 1, 25–33.
KAGEL, J. H. 1995. Auctions: A survey of experimental research. In *The Handbook of Experimental Economics*, J. H. Kagel and A. E. Roth, Eds. Princeton University Press, Princeton, NJ.
KARAT, J. 1997. Software valuation methodologies. In *Handbook of Human-Computer Interaction*, M. Helander, Ed. North-Holland, Amsterdam, The Netherlands 689–704.
KARAT, J. 2003. Beyond task completion: Evaluation of affective components of use. In *The Human-Computer Interaction Handbook*, J. A. Jacko and A. Sears, Eds. Lawrence Erlbaum, Mahwah, NJ.
LAVIE, T. AND TRACTINSKY, N. 2004. Assessing dimensions of perceived visual aesthetics of web sites. *Inter. J. Hum. Compu. Studies 60*, 3, 269–298.
LINDGAARD, G. AND DUDEK, C. 2003. What is this evasive beast we call user satisfaction? *Interacting with Computers 15*, 3, 429–452.
LINDGAARD, G., FERNANDES, G. J., DUDEK, C., AND BROWNET, J. 2006. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour and Information Technology 25*, 2, 115–126.
MEYER, J. AND SEAGULL, F. J. 1996. Where human factors meets marketing. *Ergonomics in Design 4*, 3, 22–25.
MILGROM, P. 1989. Auctions and bidding: A primer. *J. Econ. Persp. 3*, 3, 3–22.
NIELSEN, J. AND LEVY, J. 1994. Measuring usability: Preferences vs. Performance. *Comm. ACM 37*, 4, 66–75.
NORMAN, D. A. 2002. Emotion and design: attractive things work better. *Interactions 9*, 4, 36–42.
NORMAN, D. A. 2004. *Emotional Design: Why We Love (Or Hate) Everyday Things*. Basic Books, New York, NY.
ROTH A. E. AND OCKENFELS, A. 2000. Last minute bidding and the rules for ending second-price auction: Theory and evidence from a natural experiment on the Internet. *National Bureau of Economic Research*, working paper no. 7729.
SCHENKMAN, B. N. AND JONSSON, F. U. 2000. Aesthetics and preferences of web pages. *Behaviour & Information Technology 19*, 5, 367–377.
SHACKEL, B. 1991. Usability—context, framework, design and evaluation. In *Human Factors for Informatics Usability*, B. Shackel and S. Richardson, Eds. Cambridge University Press, Cambridge, UK. 21–38.
SHOGREN, J. F., SHIN, S. Y., HAYES, D., AND KLIEBENSTEIN, J. B. 1994. Resolving differences in willingness to pay and willingness to accept. *Amer. Econ. Rev. 84*, 255–270.
SU, L. T. 1998. Value of search result as a whole as the best single measure of information retrieval performance. *Inform. Proc. Manag. 34*, 5, 557–579.
TRACTINSKY, N. 1997. Aesthetics and apparent usability: Empirically cultural and methodological issues. In *Proceedings of the ACM CHI 1997 Conference on Human Factors in Computing System*. Atlanta, ACM Press New York, NY (March), 115–122.
TRACTINSKY, N. 2004. A few notes on the study of beauty in HCI. *Hum. Comput. Inter. 19*, 4, 351–357.
TRACTINSKY, N. AND LAVIE, T. 2002. Aesthetic and usability considerations in user's choice of personal media players. In *Proceedings of the 16th British HCI Conference*. London, UK (Sept.), 70–73.
TRACTINSKY, N. AND ZMIRI, D. 2006. Exploring attributes of skins as potential antecedents of emotion in HCI. In *Aesthetic Computing*, P. Fishwick, Ed. MIT Press, Cambridge, MA. 805–421.
TRACTINSKY, N., KATZ, A. S., AND IKAR, D. 2000. What is beautiful is usable. *Interacting with Computers 13*, 127–145.
VAN DER HEIJDEN, H. 2003. Factors influencing the usage of websites: The case of a generic portal in The Netherlands. *Information and Management 40*, 6, 541–549.
VICKREY, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *J. Finance 16*, 8–37.

Received April 2005; revised November 2005; accepted November 2005 by Andrew Monk