What is usable is usable

Master Thesis Information Science, Utrecht University

19 august 2004

Marijn Kampf

Supervisors
Herre van Oostendorp
Utrecht University
The Netherlands

Stephen Payne
Cardiff University
School of Psychology
Wales, United Kingdom

Thesis number INF/SCR04-21

Abstract

An increasing number of papers look at the relationship between the classic Human-Computer Interaction (HCI) interest usability and the newcomer aesthetics. We replicated Tractinsky's simulated cash machine experiment, but avoided biasing participants by showing them only one interface. We used multiple item measures and made a distinction between classic aesthetics (emphasis on orderly and clear design) and expressive aesthetics (emphasis on creativity and originality). Contrary to Tractinsky, our manipulation of usability had an effect on perceived usability and we found that classic aesthetics could influence actual usability. In the second experiment, our interfaces failed to differ in classic and expressive aesthetic judgements. We could not prove that aesthetics only had an effect on perceived usability through its effect on completion times. Expressive aesthetics appeared to affect perceived usability and not actual usability. This shows that aesthetics can influence both actual and perceived usability. It could be argued that if there is no actual usability difference, people base their perception of usability on aesthetics. If however there is an actual usability difference, people base their perception of usability on both the actual usability and aesthetics. Therefore we recommend more research into their exact relationship.

Table of content

1 Introduction

Until recently, the main interest in the Human-Computer Interaction (HCI) field has been usability related. Aesthetics was almost completely ignored in the "traditional" HCI books. Nielsen's one entry of aesthetics in the index of [NIELSEN_1993] is a reference to aesthetics, which is only mentioned in passing. In the past few years there has been an increased interest in aesthetics. Recent articles have been published with such titles as "Attractive things work better" [NORMAN_2002] and "What is beautiful is usable" [TRACTINSKY_2000]. They put forward the idea that aesthetics and usability are more closely connected than was thought. New insights into the relationship between aesthetics and usability may influence how designers and HCI experts develop interfaces.

Several experiments have been published where researchers tried to determine the relations between aesthetics and usability. In this introduction we will look at a number of articles which feature these topics.

Kurosu and Kashimura [KUROSU_1995] make a distinction between apparent usability and inherent usability. Apparent usability is an indicator for how easy people think something will be to use by looking at it and inherent usability is the usability experienced by someone actually using a system. They asked participants to rate several Automatic Teller Machine (ATM) designs on both functional and aesthetic aspects. They found a relatively high correlation between beauty and apparent usability. Furthermore, the correlation found between apparent usability and apparent beauty was higher than the correlation between apparent usability and inherent usability. Because people never actually used the devices, Kurosu and Kashimura's measurement of inherent usability is questionable. Their measure of inherent usability was based on factors that interface designers believed to enhance the inherent usability. The factors where not based on other research and they did not test whether the factors actually affected inherent usability.

Tractinsky [TRACTINSKY_1997] repeated Kurosu's et al. experiment to validate and replicate the relationships between users' perceptions of interface aesthetics and usability in a different cultural setting. He did not find the expected cultural differences between Japan and Israel, instead his results supported the basic findings by Kurosu et al. He found very high correlations between the perceived aesthetics of the interface and a priori perceived ease of use of the system. One of his conclusions is that "the results provide further support for the contention that perceptions of interface aesthetic are closely related to apparent usability and thus increase the likelihood that aesthetics may considerably affect system acceptability. … The [third] conclusion postulates that objective measures of system behaviour and use may not suffice in predicting system acceptability. Perhaps a more holistic approach toward understanding how people experience and judge information systems is needed."

Tractinsky [TRACTINSKY_2000] elaborated on these concepts in later research. He added functionality to the ATM designs by writing a computer simulation. He asked the participants about their perceptions of the system before and after they used the system. Strong correlations between the system's perceived aesthetics and it's perceived usability were found both before and after the participants used the ATM.

The aesthetics of the system affected the post-use perceptions of both aesthetics and usability, whereas the actual usability did not. He concludes: "The findings stress the importance of studying the aesthetic aspect of human–computer interaction (HCI) design and its relationships to other design dimensions."

2 Definitions

To avoid any confusion we will define a number of concepts used in this article.

2.1 Perceived vs. actual usability

Perceived usability is the user's rating of usability. It is a measure for how usable the user thinks the system is, whereas actual usability is the usability which is measured while the user uses the system. We have used the average times to complete a task as our measure of actual usability. Perceived usability was judged by the participants before (pre-usability) and after (post-usability) they used the system.

2.2 Layout vs. Interface

Tractinsky used the word *layout* in his experiments to describe the different aesthetic manipulations he made to the screen designs. We will use the more general word *interface* for two reasons. Firstly, we want to emphasise that aesthetics is more than only the layout of the buttons, images, text and other features of a single screen design. The second reason for using *interface* is that in the second experiment the elements of the interface as well as their configuration are varied.

3 Critique of Tractinsky

The experiment performed by Tractinsky in 2000 consisted of three stages. In the *pre-test stage* the ATM displayed nine different interfaces. The nine interfaces used were selected from the 26 interfaces used in Kurosu's et al. experiment [KUROSU_1995], which Tractinsky adapted to use in his replication of the experiment [TRACTINSKY_1997]. For his 2000 experiment he selected three interfaces which were rated as highly aesthetic, three which were rated low in aesthetics and three which had an aesthetic rating in between.

The experiment started by asking the participant to rate each interface on the dimensions of: aesthetics; ease of use; and amount of information on the screen. During the rating each interface was presented three times. This was followed by the test phase. The participants were assigned to one of the three aesthetic conditions; the high, low or medium aesthetic interface. Tractinsky used the participant's own scores on aesthetics to assign them to the aesthetic condition. He did this so each participant used an interface they found to have a high, low or medium aesthetic, matching the aesthetic condition they were assigned to. Participants assigned to the high condition used the screen they rated the highest and vice versa. The participants used only one screen design in the remainder of the experiment. Only after they practiced the use of the ATM were they assigned to one of two usability levels. Tractinsky used the levels low and high usability. In the low usability he added three usability difficulties. 1) Longer system delays of 9 seconds on average per task; 2) buttons that did not operate the first time they where pressed; 3) a shortcut which was not available. In the high usability condition these difficulties were not included. The participants had to perform the experimental tasks using the same interface as they used for the practise

tasks. During the post-test stage the participants were asked to rate the system on several dimensions.

We have four critiques of the way Tractinsky set up his experiment. Our first critique is that he might influence the participants by showing them multiple interfaces. With his setup, the pre-test might implicitly encourage participants to view usability only as a property that can vary between the interfaces that they compare and therefore as being a property of layout of interface objects. In contrast, participants have no reason to believe that response time would vary between devices. So when rating usability they may choose to ignore it because it does not inform the implicit comparisons they are still using.

Our second critique is that participants might try to be consistent in answering the pre- and post-experiment questionnaire. The pre-test means that any post-test judgments will have to contradict pre-test judgments if they are to show an effect of manipulated usability. The effects found could be the result of the participants wanting to appear to be consistent.

The third critique is the way in which the usability was manipulated. By adding the usability factor as an increased system delay Tractinsky perhaps does not add a salient usability factor. We think the users did not find the system delays salient because the ATM gives sufficient feedback through the display. Another possibility might be that they attribute the delays to the underlying device, for example a background or network task, instead of the ATM. Table 1 shows an overview of system response times and the effects on the user. The information in the table is based on [NIELSEN_1993]. For a realistic simulation of an ATM each task has more than one delay. It is not likely that a single delay will exceed 10 seconds. Since the users receive proper feedback they will most likely think it is a normal part of the simulation. This could explain why Tractinsky did not find an effect for actual usability. This leaves only the buttons that sometimes did not operate and the unavailable shortcut as his usability manipulation. Because the system gives proper feedback, the users will quickly notice whether their button click worked. They will still feel that the system is reacting instantaneously. They may treat it as an unavoidable constraint, nothing that could be "designed away"

Response time (approximates)	**Effects on user**
< 0.1 seconds	The user feels that system is reacting instantaneously
0.1 – 1 seconds	The user will notice the delay but his flow of thought will stay uninterrupted. The user will lose the feeling of direct manipulation.
1 – 10 seconds	The users' attention will stay on the dialogue. The user should receive feedback about the system being busy
> 10 seconds	Users will want to perform other tasks while waiting for the computer to finish.

Table 1 System response times and its effects on users Based on [NIELSEN_1993]

Our fourth critique is the questions in the questionnaire. Because so many judgements were required of the participant, Tractinsky only used single questions to measure the usability and aesthetics. We want to improve the reliability by using multiple item measures. Snijders [SNIJDERS_2003] gives two reasons why multiple item measures are superior over single item measures: improved reliability and validity. The impact of a participant not understanding a question is much smaller when using multiple item measures. The second reason is that multiple item measures are more likely to give consistent answers over time. This will give a more consistent view to the pre- and post-experimental results of the measure. Any changes found between the pre- and post-experiment measures will therefore be more likely caused by real changes in the participant's opinion than by chance. In addition to the improved reliability, using multiple item measures also enables us to improve the validity of the responses by being able to cover a broader range of the measures.

4 Research questions

The first goal of this study is to test whether the post-experimental measures still indicate strong correlations between perceived aesthetics and perceived usability if the participants see only one screen design. By showing only one screen we hope not to guide the participants in their assessment of any usability factors in the pre-experiment evaluation. Our *first* hypothesis is that post-experimental measures will not indicate strong correlations between perceived aesthetics and perceived usability if the participants see only one screen design.

The second goal of this study is to see whether a better usability manipulation does have an effect on perceived usability. Our *second* hypothesis is that a better actual usability will correspond to a better perceived usability.

The third goal of this study is to use a somewhat more elaborate measure of aesthetics and perceived usability to improve both the reliability and validity of our research.

The concepts of the two hypotheses are visualized in Figure 1.

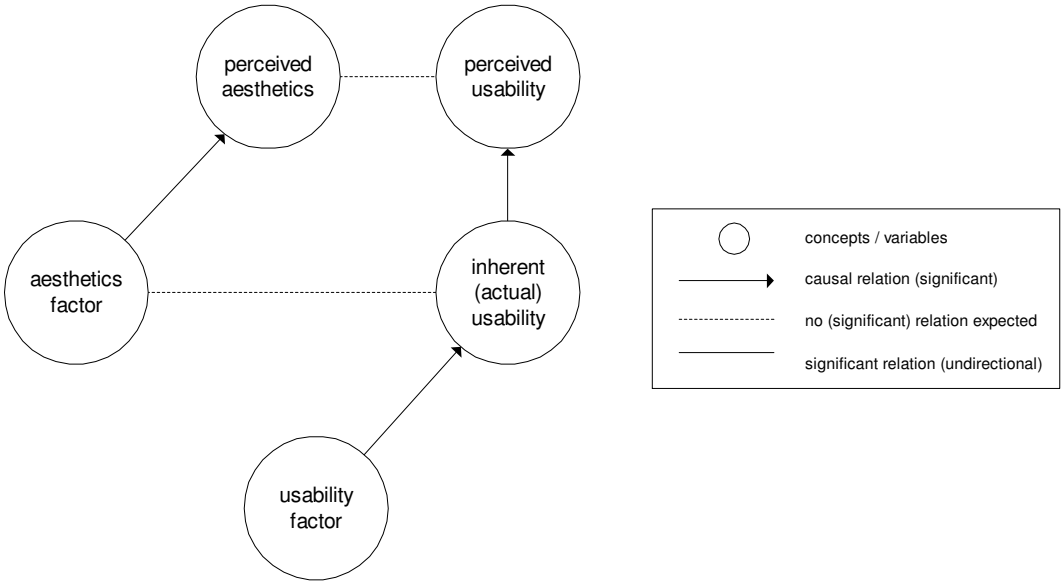

Figure 1 Visualization of the concepts and variables of the two hypothesis

5 Method

The method of this experiment is largely based on the experiment performed by Tractinsky in [TRACTINSKY_2000]. The method and the differences with Tractinsky's experiment will be outlined here.

5.1 Participants

Participants were 83 undergraduate and graduate psychology students, some of whom participated in the experiment voluntarily, some for course credit.

5.2 Design

The experiment used a 2 x 2 between groups design. Both the aesthetic and usability factors had two levels: low and high.

Aesthetics: The interfaces used for the aesthetic conditions in this experiment have been used in similar experiments before. The interfaces were introduced by Kurosu et al. [KUROSU_1995] and later adapted by Tractinsky. Tractinsky [TRACTINSKY_2000] selected nine of the 26 interfaces used by Kurosu et al. based on ratings of his earlier experiment [TRACTINSKY_1997]. He selected three interfaces for each aesthetic condition (high, low and average ratings), so he used nine interfaces in total. The two interfaces used for this experiment are the interfaces which received the highest and lowest aesthetic scores in [TRACTINSKY_2000]. The interfaces where copied as closely as possible. The only adaptation made was the deletion of the 100 and 1000 buttons, because they are not used on ATM machines in the United Kingdom. Both the high and low aesthetic interfaces consisted of the same interface objects. They differed only in the way the objects were arranged on the screen.

Participants where asked to rate the interfaces and usability on several scales. The scales are explained in further detail in the scales section below. The two interfaces used can be seen in Figure 2 and Figure 3.

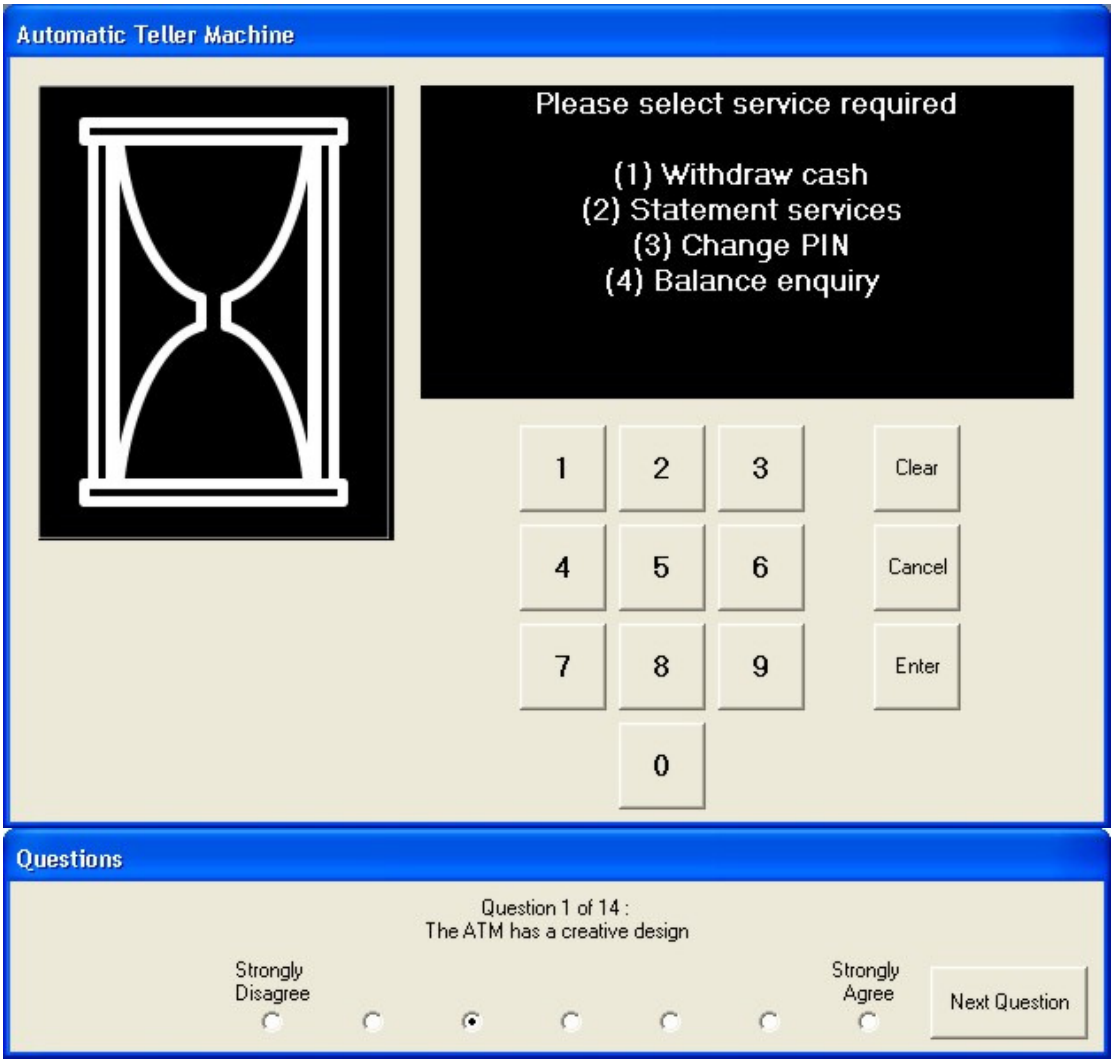

Figure 2 High aesthetics interface and window containing the pre- and post-experiment questions

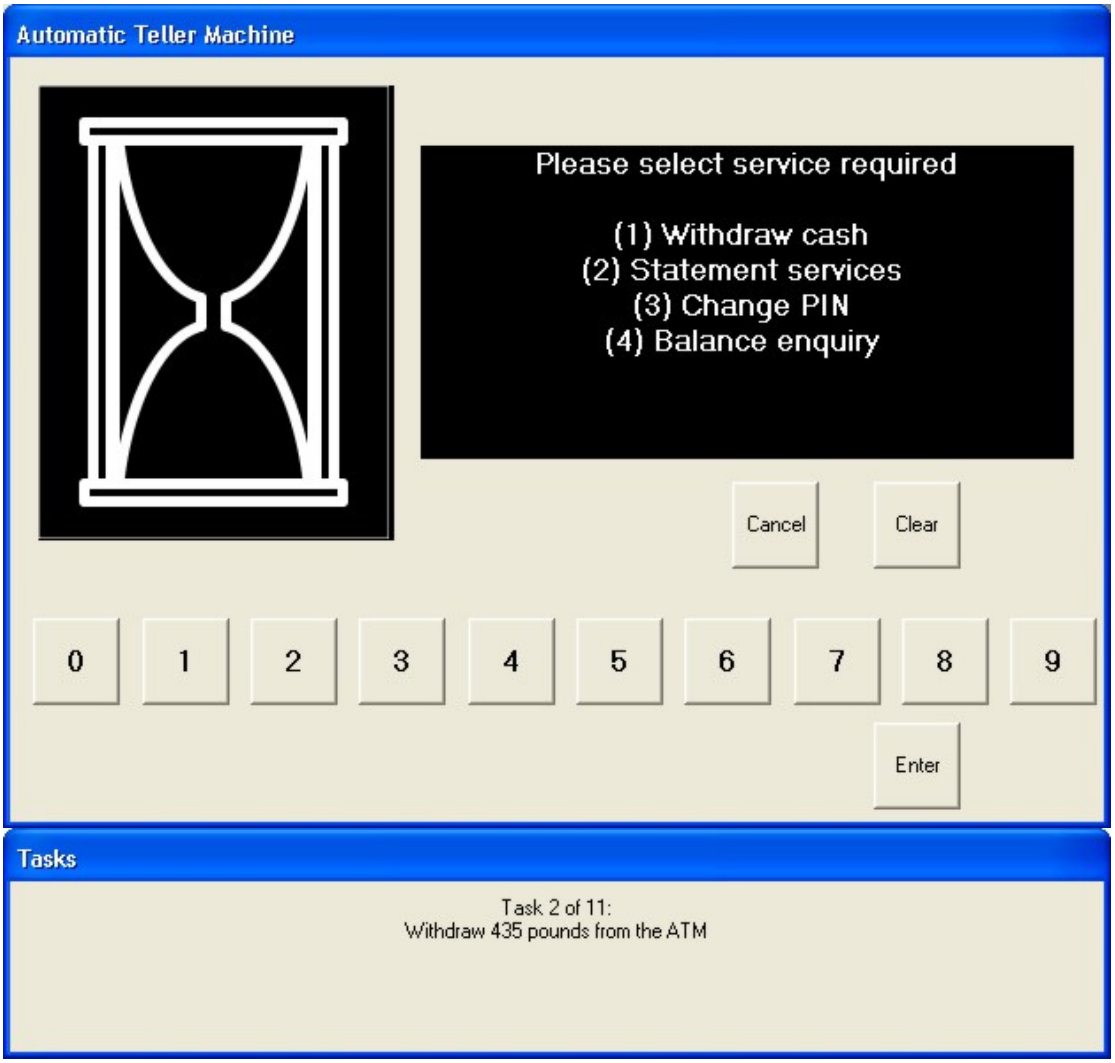

Figure 3 Low aesthetics interface and window with the tasks.

Usability: The usability factor was manipulated by adding problems in the interaction. The high usability device operated as one expects from an ATM, whereas the low usability device did not. We created the low usability condition by adding problems to the interaction in the high usability condition. The usability manipulations in the low usability condition consisted of buttons that did not always operate (randomly one out of five times a button did not function) and delays for a button to function after being clicked. A random delay between 0 and 0.5 seconds was added before the system reacted to a button click. The lag in feedback after a button click in combination with the not functioning of the buttons results in the users having to wait before knowing whether their button click worked. As in Tractinsky's experiment the participants where sensitized to the ATM's usability by asking them to complete the tasks as quickly as possible.

5.3 Procedure

Before the experiment the participants were assigned by chance to one of the four conditions. The experiment consisted of four stages. Each stage of the experiment started with a screen containing instructions for the upcoming part. The first screen explained the procedure to the participants. In the first stage the users saw the ATM

design but were unable to use it. The 14 questions about the aesthetics and usability were displayed in a separate window below the ATM. The questions were displayed one at a time in a randomized order. The participant could not proceed to the next questions without first answering the current question.
After answering the questions the participant was given four practice tasks to familiarise themselves with the ATM. Any questions the participants asked were answered either during or after the second stage.
The third stage consisted of performing the actual tasks. The tasks are described in the tasks section below. The participants were instructed to perform the 11 tasks as quickly as possible.
In the last stage the post-experiment questions were asked. These were the same questions as in the first stage. Again the questions were randomized, but extended with two questions about age and gender at the end. The experiment was concluded with a screen thanking the participant.

5.4 Tasks

The participants had to perform 11 tasks. The tasks are based on tasks used by Tractinsky but they are adapted slightly to the local situation. The tasks copied directly from Tractinsky are: *check account balance* (repeated 3 times); *withdraw cash* (repeated 4 times); *checking the account balance and withdraw cash* simultaneously (repeated 2 times); Tractinsky's last tasks was *depositing money* (repeated 2 times). Since this is not a task found on local ATM machines we replaced the tasks with *Change PIN number* and *Print a mini statement paid out*. As in Tractinsky's experiment all tasks were performed logically without exchange of physical materials. The tasks were presented one at a time in a separate window below the main ATM display. After successful completion of one task the next task was displayed automatically, the participant could only proceed to the next task after completion of the current task. The ATM display panel gave instructions on how to operate the ATM. It also indicated when the system was busy processing their requests.

5.5 Scales

To get a more refined measure of the aesthetics and usability, both dimensions were measured using multiple questions. Lavie and Tractinsky [LAVIE_2004] developed a measurement instrument of perceived web site aesthetics. They found two factors of aesthetics, "classical" and "expressive". In classical aesthetics the emphasis is on orderly and clear design, this dimension is closely related to design rules from usability experts. Creativity and originality are emphasised in the expressive aesthetics dimension. Altough the measurement was developed for websites only one element in the measure (use of special effects) was not applicable for the ATM simulation. This element was omitted in the questionnaire. The element of *Clear design* was present both in the classic aesthetic dimension and in the usability dimension. Therefore this element was omitted too. The usability dimension is based on the usability factor used in Lavie's et al. experiment and on additional usability factors based on Nielsen's book Usability Engineering [NIELSEN_1993].

The scales used comprised of:
Classic aesthetics

- The ATM has an aesthetic design
- The ATM has a pleasant design
- The ATM has a clean design
- The ATM has a symmetric design

Expressive aesthetics

- The ATM has a creative design
- The ATM has a fascinating design
- The ATM has an original design
- The ATM has a sophisticated design

Usability

- The ATM is convenient to use[1]
- The ATM is easy to operate[2]
- The ATM is easy to navigate[1]
- The ATM gives good feedback[2]
- The ATM is efficient to use[2]
- The ATM is easy to learn[2]

Usability factors based on [1]Lavie et al. or [2]Nielsen.

5.6 Variables

As in Tractinsky's experiment the variables of interest are the subjective valuations of interface properties. There are three pre-experimental variables, Aesthetics, which can be divided in classic aesthetics and expressive aesthetics, and Usability. There are several post-experimental variables where three are the same as the pre-experimental variables. The additional variables are the task times, number of buttons clicked, number of button clicks which failed and added click delays were recorded for each task.

6 Results

6.1 Outliers

Based on the average task completion times we excluded one outlier from the data. The average task completion time for low aesthetics and low usability condition was 16.5 seconds with a standard deviation of 4.4 seconds the outliers' average task completion time was 33.5 seconds.

6.2 Data

The mean ratings and standard deviations for the ATM's pre- and post-experimental variables are shown in Table 2. Each rating was scored on a scale from zero to six, except the average task times which were measured in seconds. Some additional performance indicators which were recorded by the system are shown in Table 3.

Aesthetic factor	High		Low		High		Low	
Usability factor	High		High		Low		Low	
	n = 21		N = 20		n = 21		n = 20	
	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
Pre-classic aesthetics	3.29	.78	2.53	.95	3.30	.61	2.99	.71
Pre-expressive aesthetics	1.63	.79	2.14	.96	1.81	.89	2.31	.90
Pre-perceived usability	4.41	1.00	3.93	.67	4.48	.76	4.03	.78
Post-classic aesthetics	3.54	.80	2.51	1.02	3.21	.79	2.83	.94
Post-expressive aesthetics	1.86	.99	2.40	1.00	1.99	1.19	2.24	1.11
Post-perceived usability	4.94	.94	4.43	1.05	4.46	.91	4.06	.94
Task Times (in seconds)	14.73	2.28	15.66	2.06	23.01	2.98	25.05	2.94

Table 2 Means and averages of pre and post variables

Aesthetic factor	High		Low		High		Low	
Usability factor	High		High		Low		Low	
	n = 21		N = 20		n = 21		n = 20	
	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
Added click delays (in seconds)	n/a	n/a	n/a	n/a	1.66	.17	1.73	.21
Number of failed button clicks	n/a	n/a	n/a	n/a	1.40	.43	1.44	.42
Number of buttons clicked	6.57	.41	6.36	.25	8.03	.83	8.27	1.00
Menu wait time (in seconds)	1.78	.38	1.87	.44	3.19	.74	3.08	.67

Table 3 Additional performance indicators per task

6.3 Manipulation check

A one-way analysis of variance (ANOVA) for the two aesthetic conditions (high and low aesthetic interface) revealed an effect on both of the pre aesthetics judgements. For pre-classic aesthetics the results found were ($F_{1, 80} = 5.873$; $p = 0.003$) and for pre-expressive aesthetics ($F_{1, 80} = 5.220$; $p = 0.011$).

6.3.1 Actual usability variable

The system recorded all the actions performed by the participants and the system (see Table 3). Since every participant managed to perform all the tasks successfully, we had to find a suitable indicator of actual usability. We looked at the following variables: task completion times, waiting time at the menu, number of buttons clicked per task, number of wrong menu choices, and the number of times the cancel and clear buttons were clicked. One-way analysis of variance (ANOVA) for the usability condition showed a significant result on the task completion times ($F_{1, 80} = 220.776$; $p < .001$), menu waiting time ($F_{1, 80} = 106.907$; $p < .001$) and number of buttons clicked per task ($F_{1, 80} = 121.397$; $p < .001$). The three variables were highly correlated with each other, see Table 4.

	Waiting time at menu	Number buttons clicked per task
Task completion time	.85	.82
Waiting time at menu		.64

Table 4 Correlation table of actual usability variables. The correlations are significant at the .01 level (2-tailed)

Because the correlation between task completion time and the other two variables was the strongest (>= .82) we decided to use the average completion times of the tasks as our variable for actual usability.

The usability manipulation was evaluated by comparing the average completion time of the 11 tasks for both usability conditions. A two-way ANOVA was used to find the effects of the usability and aesthetics factors on completion times. The usability factor ($F_{1, 78} = 236.975$; $p < 0.001$) had a significant effect on the completion times. Contrary to Tractinsky we also found a significant effect of the aesthetics interface factor on completion times ($F_{1, 78} = 6.725$; $p = .011$). We found no interaction effect between usability and aesthetics.

The average completion time in the high usability (no delay) conditions was 15.2 seconds, whereas the average for the low usability conditions was 24 seconds. Of the 8.8 seconds difference, 1.7 seconds was caused by the added click delays in the buttons. Of the remaining 7.1 seconds the users in the low usability conditions waited 1.3 seconds longer at the menu of the ATM before continuing with the next task. The additional delays for each condition are shown in Table 3. The other 5.8 seconds must be attributed to the other usability differences added to the system. This indicates that the usability manipulation succeeded.
The average completion time with the high aesthetics interface was 18.9 seconds whereas the average completion time in the low aesthetics condition was 20.4 seconds. There is a 1.5 seconds difference between the two aesthetic conditions.

6.4 Aesthetics vs. Usability

6.4.1 Correlation analysis

For comparison with Tractinsky, the correlations among the perceived measures are reported in Table 5.

	Pre-expressive aesthetics	Pre-perceived usability	Post-classic aesthetics	Post-expressive aesthetics	Post-perceived usability	Task times
Pre-classic aesthetics	.186	.346(**)	.718(**)	.269(*)	.431(**)	.034
Pre-expressive aesthetics		-.022	.339(**)	.796(**)	.159	.132
Pre-perceived usability			.250(*)	.048	.473(**)	-.009
Post-classic aesthetics				.460(**)	.621(**)	-.018
Post-expressive aesthetics					.311(**)	.017
Post-perceived usability						-.287(**)
** Correlation is significant at the .01 level 2-tailed.						
* Correlation is significant at the .05 level 2-tailed.						

Table 5 Correlation matrix of pre, and post-experimental measures (n = 82)

The perception of pre-classic aesthetics and pre-perceived usability were correlated (r = .346). There was not a significant correlation between expressive aesthetics and perceived usability before the use of the ATM. The correlation of classic aesthetics resembles the results found by both Kurosu and Kashimura [KUROSU_1995] and by Tractinsky [TRACTINSKY_1997, TRACTINSKY_2000], although the correlation found is lower. The correlations between the two aspects of aesthetics and perceived usability increase after the system is used. The correlation between post-classic aesthetics and post-perceived usability (r = .621) increases and after the experiment there is a correlation between post-expressive aesthetics and perceived usability (r = .311). The task times correlated only with the measure of post-perceived usability (r = -.287). While there is not a correlation between the pre-classic aesthetics and pre-expressive aesthetics, there is a correlation between post-classic aesthetics and post-expressive aesthetics (r = .460). The correlation for the pre and post ratings of aesthetics is very high: classic aesthetics (r = .718) and expressive aesthetics (r = .796). The much lower correlation between pre- and post-perceived usability (r = .473) indicates that while giving their opinions after the experiment, the participants did not replicate the answers they gave before experiment.

6.4.2 Regression analysis

A regression analysis was performed to determine whether the perceived post-usability was influenced by the interface after the effect of the actual usability was removed by taking out the average task times. The dependent variable used in the regression analysis was the average of post-experiment perceived usability. The independent variables used were perceived pre-usability, pre-classic aesthetics, pre-expressive aesthetics, post-classic aesthetics, post-expressive aesthetics, task times and both the usability and interface factors as independent variables. The regression described 51.8% of the variance (R^2_{adj} = 49.9%) and the overall relationship was significant ($F_{3, 78}$ = 27.923; $p < .001$). The judgements of post-classic aesthetics (t_{78}=6.471; beta = .529; $p < .001$), perceived pre-usability (t_{78} = 3.522; beta = .288; p = .001) and task times (t_{78} = -3.271; beta = -.257; p = .002) were significant predictors. The other independent variables were excluded by the stepwise regression.

6.4.3 The interface and actual usability

To see whether the added delay influenced the effect the interface had on actual usability, we analyzed the simple main effects, even though the interaction effect of aesthetics and usability delays was not significant. The aesthetic interface condition influenced the average task times in the added delay condition ($F_{1, 78} = 6.334$; $p = .014$), but not in the no delay condition ($F_{1, 78} = 1.324$; $p = .253$). In other words the additional actual usability effect of the interface factor only occurred in the cases where there already was a usability delay.

6.4.4 Layout and perceived usability

An analysis of variance was performed for perceived post-usability. A two-way ANOVA was used to find the effects of usability and aesthetics factors on perceived post-usability. Both the usability factor ($F_{1, 78} = 4.017$; $p = .049$) and the aesthetic factor ($F_{1, 78} = 4.542$; $p = .036$) were significant. There was no significant interaction.

6.5 Visualization

In Figure 1 we gave a visual representation of the hypothesis. In Figure 4 a summary of the concepts and variables found in the results are visualized in a similar manner. To keep the visualization readable only the most important relations between the pre and post aesthetic concepts are visualized.

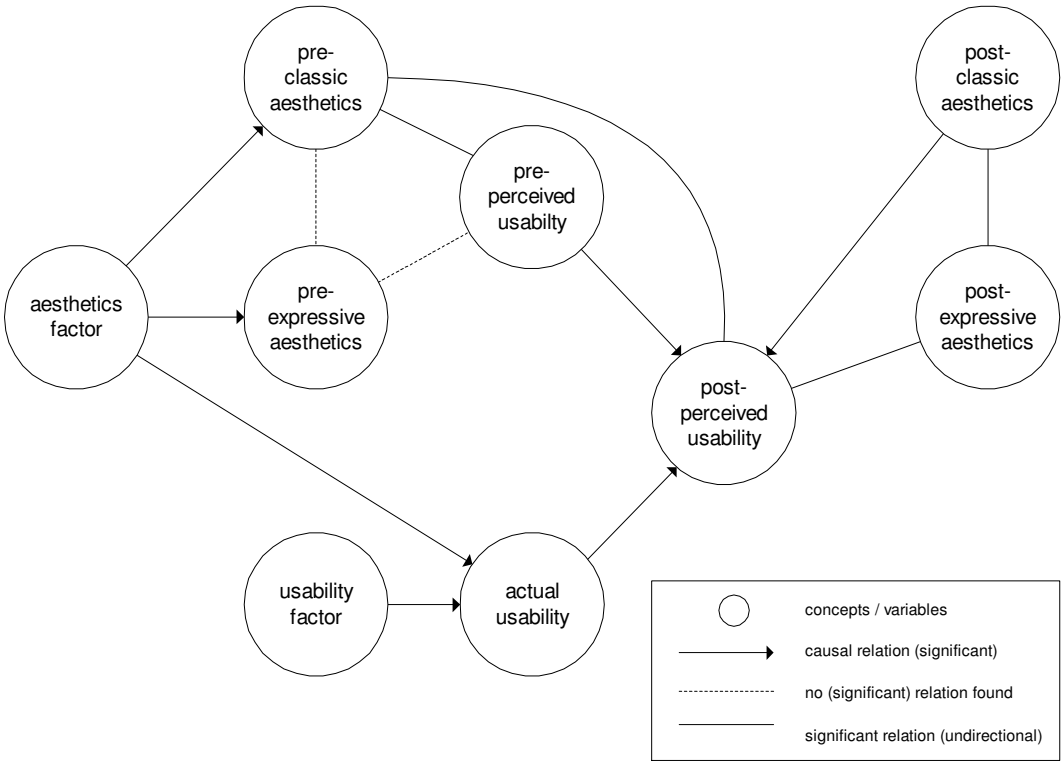

Figure 4 Visualization of the results of the concepts and variables

7 Discussion

The main part of the discussion will focus on comparing the results of this experiment to the results of Tractinsky's [TRACTINSKY_2000] experiment.

In Tractinsky's experiment the participants where asked to rate several different screen designs in the pre-experiment phase. We believe that this guided the participants in their assessment of the usability factors. We therefore showed each participant only one screen design during the entire experiment. The first goal of this study was to determine whether the post-experimental measures still indicate strong correlations between perceived aesthetics and perceived usability when we use only one design per participant.

We did find strong correlations between the post-experimental measures of aesthetics and perceived usability. Therefore we had to reject our *first* hypothesis. However, contrary to Tractinsky's study, we found an effect of the interface on actual usability of the system. This effect of the interface could also explain the correlation we found between classic aesthetics and perceived usability.

The second goal of this study was to see whether a different manipulation of the usability factor had an effect on the perceived usability as well as on the actual usability. We did find a significant beta (-.257) between actual usability on perceived usability. We can therefore accept our *second* hypothesis. Tractinsky did not find such an effect even though the delays we added (on average 1.7 seconds per task) where much shorter than his (on average 9 seconds per task). The place where the delays were added might explain this. Tractinsky's manipulation of the usability consisted of added system delays, buttons that did not always operate the first time they were pressed and a shortcut which could not be used. We also had buttons which did not always operate. The delays in our experiment were not added as additional system delays, but as a delay after a button was clicked. If the user clicked on a button in the low usability version, a random delay (on average 0.25 seconds per click) was added before the button click was processed. Tractinsky's usability manipulation might have been less salient to the users. This could have been because the system gave the proper feedback by displaying a busy hourglass in both conditions. The only difference between his two groups was the slight increase in the time the participants had to wait. The added delay was in a place where the users expected the system to be busy anyway. Our usability manipulation was in a place where the participants did not expect it in a properly functioning system. At delays longer than 0.1 seconds users stop feeling that the system is reacting instantaneously [NIELSEN_1993]. We feel that because the users noticed the usability manipulation they also reported it in their judgements of usability. Furthermore, according to our analysis, the participants were implicitly guided in the first phase of Tractinsky's experiment to treat "usability" as a property inherent in the layout of the device.

The simple effects analysis showed that the effect of aesthetics on actual usability was only true in the delay condition and not in the no-delay condition. The second simple effect analysis showed that the effect of aesthetics on perceived usability was only true in the no-delay condition and not in the delay condition. The regression analysis showed an effect for post-classic aesthetics but not for pre-classic aesthetics. This

raises the idea that aesthetics only has an effect on perceived usability through its effect on completion times.

The third goal of this study was to use more elaborate measures of aesthetics and perceived usability. Using this measure we were able to make a distinction between classic and expressive aesthetics. This made it possible to give a more detailed view of the aesthetics factor. We found a difference between the two types of aesthetics. The pre-experiment classic and expressive aesthetics were not correlated. The post-experiment classic and expressive aesthetics on the other hand did correlate. Thus, using the system made a difference to the way the participants rated the aesthetics of the device. There was no correlation found between pre-experiment expressive aesthetics and perceived usability, however the post-expressive aesthetic and perceived usability did show a correlation.

8 Conclusions

This study showed that the actual usability does have an effect on perceived usability if the usability is manipulated in a way that it influences the user. Additionally, and in keeping with Tractinsky's findings, judged aesthetics also had an effect on perceived usability. This makes way for the notion that the aesthetics effect on perceived usability may have been due to the effect on actual usability.

9 Experiment II

In the first experiment the idea was formed that aesthetics only has an effect on perceived usability through its effect on completion times. In the second experiment we want to further test for an effect of aesthetics on perceived usability in a situation where we do not expect an effect on actual usability.

10 Research questions

The first goal of this experiment is to test whether aesthetics influences perceived usability if there is no effect of aesthetics on actual usability. Our first hypothesis is that aesthetics does not influence the perceived usability if aesthetics do not influence actual usability. The hypothesis is visualized in Figure 5.

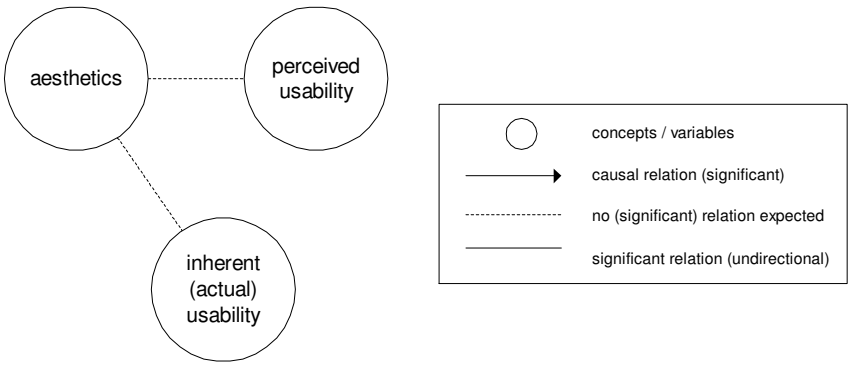

Figure 5 Visualization of the concepts and variables of the hypothesis

The second goal of this experiment is to test whether we can make designs which differ on the classic and expressive aesthetic dimensions and see if the two dimensions have a different influence on perceived usability.

12 Method

The method of the second experiment is similar to the method of the first experiment. Only the differences between the two experiments are listed here.

12.1 Participants

Participants were 57 undergraduate and graduate psychology students who participated in the experiment either voluntarily, or for course credit. There were an additional 26 participants, mainly students who participated in the experiment by completing a downloadable version of the experiment on their own computer. The results of the 83 participants were pooled for the data analysis.

12.2 Design

The experiment used a 2 x 2 between groups design. Both the classic aesthetic factor and expressive aesthetic factor had two levels: low and high.

Classic aesthetics: The results from the first experiment showed that our manipulation of aesthetics was only significant for the classic aesthetic conditions. We therefore use the aesthetic factor from the first experiment as the classic aesthetics factor in the second experiment.

Expressive aesthetics: The expressive aesthetics factor of the ATM was created by improving the looks of the interface for the high expressive aesthetic condition and by deteriorating the looks for the low expressive aesthetic condition.

The four ATM designs are depicted in Figure 6.

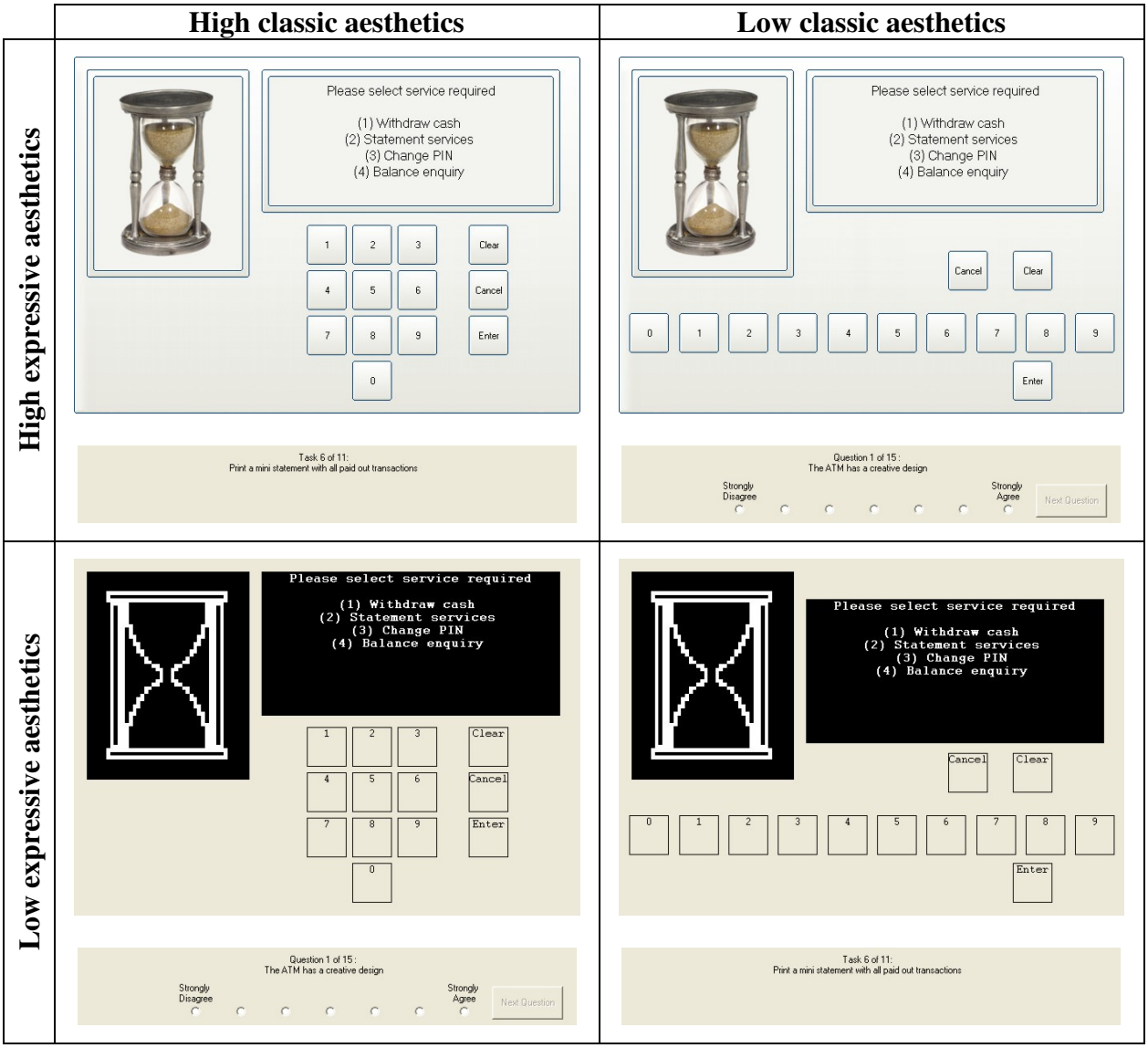

Figure 6 The four classic and expressive aesthetic ATM designs

12.3 Procedure, tasks and scales

The procedure, tasks and scales where almost identical to the first experiment. The only change made was the addition of two questions to improve the exact comparison to Tractinsky's experiment. The question "The ATM is easy to use" was added to the pre- and post-questionnaire and a question to measure satisfaction was added to the post-questionnaire. The second experiment had 15 pre-experiment and 18 post-experiment questions.

12.4 ATM screen size

For the second experiment the same computer was used as in the first experiment. In the second experiment the ATM machine and the questions were displayed larger than in the first experiment. This was because the computer was not able to display the required number of colours for the high expressive factor. The resolution of the monitor was increased, this resulted in the ATM taking up a bigger percentage of the screen than in the first experiment. Therefore the expectation is an increase in the usability effect of the classic aesthetic condition.

12.5 Variables

The variables of interest are the same as in the first experiment with specific attention to both aesthetic factors.

13 Results

13.1 Data

The mean ratings and standard deviations for the ATM's pre- and post-experimental variables are shown in Table 6.

Expressive aesthetics factor	**High**		**High**		**Low**		**Low**	
Classic aesthetics factor	**High**		**Low**		**High**		**Low**	
	(n = 19)		**(n = 20)**		**(n = 23)**		**(n = 21)**	
	Mean	**St.dev.**	**Mean**	**St.dev.**	**Mean**	**St.dev.**	**Mean**	**St.dev.**
Pre-classic aesthetics	3.36	1.03	2.94	1.15	2.97	.95	2.43	.82
Pre-expressive aesthetics	2.04	1.10	2.43	1.33	1.84	1.10	1.83	.89
Pre-perceived usability	4.17	.89	3.87	1.06	4.38	.96	4.24	.97
Post-classic aesthetics	3.33	.93	3.00	1.18	3.07	.90	2.42	.79
Post-expressive aesthetics	2.24	1.19	2.56	1.21	2.05	1.44	2.11	1.01
Post-perceived usability	4.60	.96	4.21	1.06	4.93	.73	4.71	.74
Task times (in seconds)	16.03	3.32	14.93	2.25	14.56	2.95	15.40	2.46
Menu wait time (in seconds)	2.04	.66	1.96	.64	1.82	.51	1.66	.36
Number of buttons clicked	6.64	.53	6.40	.38	6.44	.45	6.59	.45

Table 6 Means and averages of pre and post variables of the second experiment

13.2 Manipulation check

The usability manipulation was evaluated by comparing the average completion time of the 11 tasks for both the classic and expressive aesthetic interface conditions. A two-way ANOVA was used to find the effects of the two types of aesthetics. Neither factor nor the interaction had a significant effect.

We evaluated our manipulation of the two aesthetic factors by comparing the pre-experimental judgements for classic and expressive aesthetics for both aesthetic conditions using a two-way ANOVA. The results showed a significant effect on judged classic aesthetics of both the classic ($F_{1, 79}$; 4.833; p = .031) and expressive aesthetics ($F_{1, 79}$; 4.248; p = .043) factors. There was no effect on judged expressive aesthetics. No interaction effect was found in either test. These results indicate that we succeeded in our manipulation of classic aesthetics but failed to manipulate the expressive aesthetics.

13.2.1 Correlation analysis

Correlations between the perceived measures both before and after the participants used the system are presented in Table 7. The correlations of which the significance changed compared to the first experiment are printed in italics.

	Pre-expressive aesthetics	Pre perceived usability	Post-classic aesthetics	Post-expressive aesthetics	Post-perceived usability	Task times
Pre-classic aesthetics	*.406(**)*	.360(**)	.809(**)	.358(**)	.456(**)	*.299(**)*
Pre-expressive aesthetics		.034	.397(**)	.755(**)	.047	.006
Pre-perceived usability			.246(*)	.031	.540(**)	.124
Post-classic aesthetics				.475(**)	.543(**)	*.281(**)*
Post-expressive aesthetics					*.215*	.026
Post-perceived usability						*.028*
** Correlation is significant at the .01 level 2-tailed.						
* Correlation is significant at the .05 level 2-tailed.						

Table 7 Correlation matrix of pre, and post-experimental measures (n = 83)

We now find a fairly strong correlation between pre-classic aesthetics and pre-expressive aesthetics. While there was a correlation between post-expressive aesthetics and post-perceived usability in the first experiment, using the second version of the ATM does not cause the correlation to appear. Furthermore now both pre- and post-classic aesthetics correlate with task times and we did not find a correlation between post-perceived usability and task times.

13.2.2 Perceived usability

The effects of the classic and expressive aesthetic factors on perceived post-usability were tested using a two-way ANOVA. We found a significant result only for expressive aesthetics ($F_{1, 79} = 4.433$; $p = .038$). However, the average score in the low expressive aesthetics design (4.8) was higher than the average score in the high expressive aesthetic design (4.4).

We compared the average time the participants waited at the ATM menu for both the classic and expressive aesthetic interface conditions. A two-way ANOVA was used to find the effects of the two types of aesthetics. The expressive aesthetic condition was significant ($F_{1, 79} = 4.698$; $p = .033$) whereas the classic aesthetic condition was not.

13.2.3 Regression analysis

A regression analyses was performed to determine the variables which influence perceived post-usability. The dependent variable used in the regression analyse was the average of post-experiment perceived usability. The independent variables used were perceived pre-usability, pre-classic aesthetics, pre-expressive aesthetics, post-classic aesthetics, post-expressive aesthetics, task times and both the usability and interface factors as independent variables. The regression described 61.3% of the variance ($R^2_{adj} = 58.8\%$) and the overall relationship was significant ($F_{5, 77} = 24.407$; $p < .001$). The results of the significant independent variables are shown in Table 8.

	t_{77}	beta	p
Pre-perceived Usability	5.710	.437	.000
Post-classic aesthetics	5.402	.480	.000
Pre-expressive aesthetics	-4.000	-.439	.000
Expressive aesthetics factor	-3.012	-.225	.004
Post-expressive aesthetics	2.316	.270	.023

Table 8 Results of significant independent variables

13.3 Visualization

As in the first experiment Figure 7 shows the visualization of the results. To keep the visualization readable only the most important relations between the pre and post aesthetic concepts are visualized.

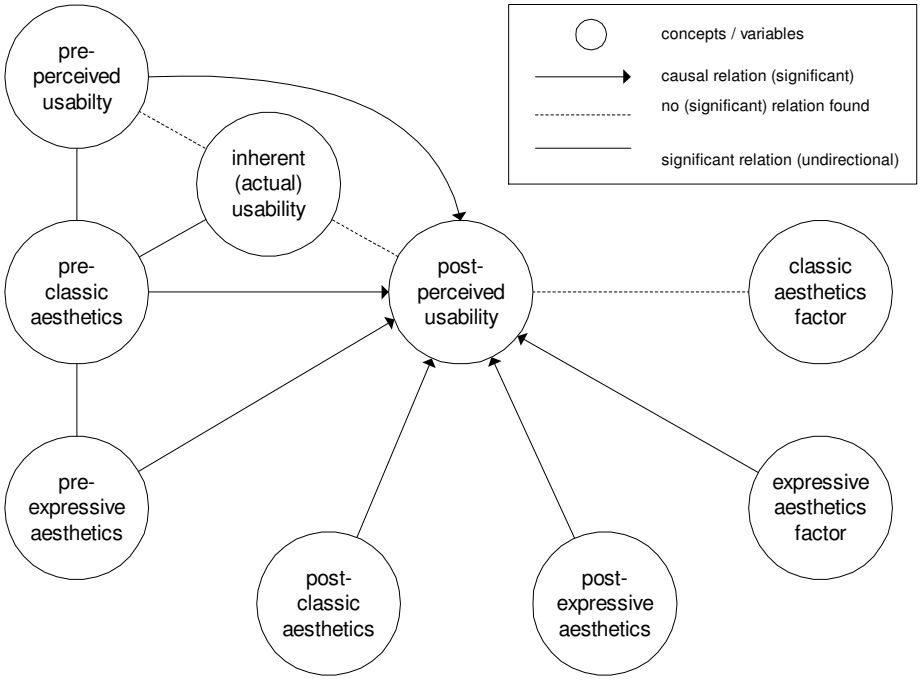

Figure 7 Visualization of the results of the concepts and variables

13.4 Discussion

Our hypothesis was that aesthetics will not influence the perceived usability if they do not influence actual usability. We have to reject this hypothesis since we found an effect of expressive aesthetics on perceived usability, while there was no effect of expressive aesthetics on actual usability. We also found a correlation between both pre- and post-classic aesthetics and task times while we did not find a correlation between post-perceived usability and actual usability. We did find an effect of expressive aesthetics on the time the participants waited in the menu. A possible explanation for this might be that the changes we made to the colours and font of the display decreased the readability.

The second goal of this experiment was to test whether we can make designs that differ on the classic and expressive aesthetic dimensions and see if the two dimensions have a different influence on perceived usability. We failed to create interfaces which differ in the expressive aesthetics of the interface. There are two

possible explanations for this. The first is that, despite our intention, the alterations we made were changes to the classic aesthetics instead of the expressive aesthetics. A second explanation is that the participants do not associate expressive aesthetics with an ATM machine or that the difference between classic and expressive aesthetics is too small to notice when evaluating our ATM machine.

14 General discussion

The results of our first experiment contradicted the results of Tractinsky, while our second experiment did not provide the results we expected and were more similar to Tractinsky's findings. It could be argued that if there is no actual usability difference and that people base their perception of usability on aesthetics. However, if there is an actual usability difference people base their perception of usability on both the actual usability and on aesthetics.

We think the differences in the two experiments might be caused by the usability manipulation. In our first experiment we manipulated the usability in such a way that it had an effect on actual usability. This caused the actual usability to be the largest influence on perceived usability. Due to the lack of the usability manipulation, perceived usability in the second experiment was based on pre-perceived usability and the classic and expressive aesthetics of the ATM.

Tractinsky argued "What is beautiful is usable" based on his finding that his manipulation of aesthetics showed an effect on perceived usability and not on actual usability. In the first experiment however we found that aesthetics did have an effect on actual usability. Furthermore we found that actual usability had an effect on perceived usability. Unsurprisingly we therefore also found an effect from the aesthetic judgements on perceived usability. Therefore it is possible to argue "What is usable is usable".

Our findings are based on the results of a limited experiment. We were only able to use a relatively small number of participants. Most of our participants were psychology undergraduates, it could be that the same experiment performed by for example older people or people with another profession would have provided different results. Park, Choi and Kim [PARK_2004] give a short overview of literature on differences for various demographic factors in aesthetic perceptions. They mention the demographics of age, gender, education levels, levels of expertise, prior experiences, religious background and lifestyle.

We looked at the influence of classic and expressive aesthetics while participants used a simulated ATM machine. In general, aesthetics are not highly associated with ATM machines. The question remains as to whether our results could be generalised to apply to other applications or products.

In the first experiment we set out to manipulate the aesthetics of the ATM interface. We had no preset intention to specifically manipulate the classic or expressive aesthetic dimension. According to the aesthetic judgements of the participants we only manipulated the classic aesthetics of the interface. The changes of the second experiment were intended to be changes to the expressive aesthetics of the interface. Despite our intentions the participants' judgements only differed for the classic

aesthetics and not for the expressive aesthetics. This raises a question about what specific elements and features of screen designs contribute to the judgements of classic and expressive aesthetics.

We mainly looked at the functional part of the ATM; the interface with the feedback screens and the input buttons. It could be argued that aesthetics of a product can be divided into two parts. The aesthetics of the part used in operating the product and the aesthetics of the remainder; the box. An example to differentiate between the two is an iMac computer. The physical computer is a highly stylised object, which would be the box, while its operating system, Mac OS, and the peripheral equipment such as its keyboard would be the functioning part. It would theoretically be possible to create a PC running e.g. Windows XP in the box of the iMac. Based on our findings we would expect that the aesthetics of the box would have no influence on usability whereas the aesthetics of the operating system could through its influence on actual usability. We found only an influence of aesthetics on usability if there was no actual usability effect. Due to the actual usability problems caused by the complexity of an operating system we expect only no influence of "the box".

The dictum "form follows function" has had several periods of followers in architecture. In architecture it has lost its appeal due to a limited interpretation of function [BANHAM_1972]. We think the dictum could apply to the functional part of designs. Based on Tractinsky's title we summarised the findings of our research with the title "What is usable is usable", indicating that aesthetics influences perceived usability due to its influence on actual usability. We could also say that the form is therefore dependent on the function.

15 General conclusions

Our experiment showed that aesthetics could influence both actual and perceived usability. This implicates that designers of interfaces should not only pay attention to the usability of their designs but also to the aesthetics of their designs.

16 Acknowledgements

We would like to thank Tractinsky for his quick reaction to our inquiry and letting us use his screen designs.

17 References

[BANHAM_1972]	Reyner Banham; Theory and design in the first machine age; 1972; Pages 320-321
[KUROSU_1995]	Masaaki Kurosu and Kaori Kashimura; Apparent Usability vs. Inherent Usability Experimental analysis on the determinants of the apparent usability; Association for Computing Machinery; CHI 95 - Conference on Human Factors in Computing Systems; 1995
[LAVIE_2004]	Talia Lavie and Noam Tractinsky; Assessing dimensions of perceived visual aesthetics of web sites*1; International Journal of Human-Computer Studies; Volume 60; Issue 3; March 2004; Pages 269-298

[LINDGAARD_2003]	Gitte Lindgaard and Cathy Dudek; What is this evasive beast we call user satisfaction?; Interacting with Computers; Volume 15; Issue 3; June 2003; Pages 429-445
[NIELSEN_1993]	Jakob Nielsen; Usability Engineering; Morgan Kaufmann; 1993;
[NORMAN_2002]	Don A. Norman; Emotion and design: Atrractive things work better; Interactions Magazine; Volume 4; July/August 2002; Pages 36-42; http://www.jnd.org/dn.mss/Emotion-and-design.html
[NORMAN_2003]	Don A. Norman; Emotional Design: Why We Love (Or Hate) Everyday Things; Basic Books; 23-Dec-03; Chapter 1; http://www.jnd.org/ED_Draft/CH01.pdf
[PARK_2004]	Su-e Park, Dongsung Choi and Jinwoo Kim; Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users; Interacting with Computers; Volume 16; Issue 2; April 2004; Pages 351-376
[SCHAIK_2004]	Paul van Schaik and Jonathan Ling; The effects of screen ratio and order on information retrieval in web pages; Displays; In Press, Corrected Proof; Available online 10 February 2004;
[SNIJDERS_2003]	Tom A.B. Snijders; Multilevel Analysis in M. Lewis-Beck, A.E. Bryman, and T.F. Liao (eds.); The SAGE Encyclopedia of Social Science Research Methods; Volume 2; 2003; Pages 673-677; http://stat.gamma.rug.nl/snijders/MultilevelAnalysis.pdf
[TRACTINSKY_1997]	Noam Tractinsky; Aesthetics and Apparent Usability: Empirically Assessing Cultural and Methodological Issues; Association for Computing Machinery; CHI 97 - Conference on Human Factors in Computing Systems; 1997;
[TRACTINSKY_2000]	N. Tractinsky, A. S. Katz and D. Ikar; What is beautiful is usable; Interacting with Computers; Volume 13; Issue 2; December 2000; Pages 127-145