applied sciences

MDPI

Article

A Comparative Study of Design Evaluation with Virtual Prototypes Versus a Physical Product

Chih-Hsing Chu [1,*] and Erh-Ting Kao [2]

[1] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 30013, Taiwan
[2] AsusTek Computer Inc., Taipei 11259, Taiwan; cjliao@ie.nthu.edu.tw
* Correspondence: chchu@ie.nthu.edu.tw

Received: 14 May 2020; Accepted: 6 July 2020; Published: 9 July 2020

Abstract: Design evaluation is an important stage in the product development process. Virtual prototypes enable economic design evaluation with higher flexibility, but the evaluation effectiveness may be limited compared to that of the real product. Few studies have analyzed whether or not virtual prototypes are comparable with the real product on the evaluation of product attributes. In this study, we conducted two-stage experiments to compare the effectiveness of design evaluation by using virtual prototypes versus the product they aim to represent. Numerous design features were evaluated from a physical appearance and usability point of view with assessment measurements including performance accuracy and the emotional responses of the users. The experimental results revealed that the visual virtual prototypes were typically not as effective in estimating the physical and appearance features, while no significant difference was observed in the usability between the evaluation media. The visual virtual prototypes tended to invoke more negative and passive emotional states in comparison to the actual product. However, with the addition of instant sensory feedback, the emotional responses were raised to a more positive and active level, which was similar to the one observed with the physical product. The findings of this study indicate the shortcomings of using virtual prototypes in the design evaluation process. Our conclusions may assist future studies in improving the practicality of virtual prototyping by the addition of useful features.

Keywords: design evaluation; virtual prototypes; emotional responses; usability; user experience

1. Introduction

Design evaluation is an essential phase in the product development process. This is particularly true when a company plans to develop and release a new product. In the conventional approach, physical prototypes are produced for the purpose of verifying the design functions and collecting user feedback. In different stages of the design process, the prototypes' level of realism may vary depending on use. The lifecycle of current consumer products is relatively short. Using physical models to verify the product design may no longer be an effective solution, considering the cost, time, and effort required to produce physical prototypes [1]. Advances in computer technologies allow for a more realistic simulation, i.e., developing virtual prototypes of the final product. Although a visual virtual prototype has various obvious limitations in terms of design evaluation, e.g., haptic or textile feedback is difficult to realize with current rendering technologies, it can still adequately simulate the visual aspects of a product [2]. Design features, such as colors or textures, can be replaced and assessed visually, and require a short amount of rendering time. This may explain why most previous studies focused on the visual aspects of design evaluation by using virtual prototypes [3].

Early researchers such as [4] argued that a good product could fulfil the user's needs at the functional, usability, and emotional levels. They suggested that more comprehensive approaches,

such as those including emotional aspects, should be introduced in the design evaluation stage. Söderman [5] investigated how sketches, physical models, and virtual media contribute to the evaluation of card designs. It was found that the familiarity of the product influenced the evaluation result, even though the physical models were still perceived to be more realistic. Ibrahim and Rahimian [6] found that hand sketches could communicate the design concepts more intuitively. They reported that computer aided media impose somewhat of a constraint to the designer's creativity. They compared 2D virtual prototypes with the actual products and found the former to be effective with regard to evaluation of simple design features.

Gibson et al. [7] proposed two different experimental implementations for the real-time integration of virtual and physical prototyping based on computer aided design (CAD) techniques and rapid prototyping. A digital mock-up (DMU) was developed for use in an experiment that compared simulated assembly tasks in both real and virtual environments [8]. The subjective evaluation results of the real (RE), virtual (VE), and virtual with force feedback (VEF) environments indicated a real sensory and difficulty gap between the RE and VEF, while a smaller difference was observed between the RE and the VE. Kim et al. [9] analyzed user impressions of a product using virtual prototyping. They characterized the relationship between user impressions and design elements of automobile interior through experiments. Experimental data validated that virtual prototyping can help analyze user impressions of design alternatives at the early stage of the design process. Aromaa and Väänänen [10] conducted an experimental study for comparing augmented reality (AR) and virtual environment (VE) prototypes of a rock crushing machine. The experimental result indicated that the VE system was more suitable to support the assessment of visibility, reach, and the use of tools than the AR system. The previous study [11] emphasized the importance of working prototypes in product development and the importance of obtaining the users' reactions upon interacting with such prototypes. They demonstrated how virtual reality (VR) prototypes do not only represent a valid alternative to physical prototypes, but also take a step forward owing to the possibility of simulating multisensory and real-time modifiable interactions between the user and the prototypes. Faust et al. [12] conducted a preliminary experimental study of mixed reality prototyping for its ability to be used to evaluate usability and user experience aspects of a real projector. Both mixed reality prototype and real product showed a longer time of use and more errors in the use as the difficulty of the task increased. The result of user experience evaluation was comparable for both.

The form and type of virtual model representation will influence human assessment in various design disciplines such as product design [13], interior design [14], and landscape design [15]. Bligård et al. [16] compared two physical prototypes of different scales and a CAD model representing a ship bridge workstation. Participants were asked to assess the proposed design and to compare the models' relative merits. The physical models received more positive feedback than the CAD model, both regarding content richness and quantity. In the experimental study of Voit et al. [17], subjects compared five different methods (online, virtual reality, augmented reality, lab setup, and in situ) to evaluate early prototypes of smart artifacts using different standardized questionnaires. The experimental results revealed that evaluation methods significantly influence the assessment result. This implies that results may not be compared across studies that use different methods even using standardized questionnaires. Kuliga et al. [18] compared the human–environment interaction between a real building and a virtual model of the same building in VR environment. Both quantitative (bipolar semantic differential questions) and qualitative (interview) measures were collected from experiments. They found few differences in the quantitative analysis result, but significant qualitative differences. Additionally, focused on simulated environments, Higuera-Trujillo et al. [19] compared the psychological and physiological responses evoked by photographs, 360° panoramas, and virtual reality against a physical environment. They found that the 360° panorama is the most valid display format according to psychological responses and virtual reality is the most valid one in terms of physiological response.

This study is different than the existing research in the following regards. First, we proposed an experimental procedure that systematically compares the virtual prototype with a real product on design evaluation from a holistic view. The evaluation was carried out from three aspects: Functional, usability, and emotional response by the measure of single-answer questions, procedural correctness, and an emotional scale, respectively. Previous studies conducted a comparative experiment on virtual and physical prototypes either from single [5,7,8,10,20] or two aspects [12,19]. Second, most of them [16–19] have treated the product itself as a golden standard and assessed the similarity between the product and its virtual counterpart. Although this approach may sound natural, we argue that the evaluation should aim beyond comparing the prototypes against an existing target object. In this study, we used external references instead of the product itself as the evaluation object. These references have definite values derived from the physical product. In addition, some studies on design evaluation with virtual prototypes focused on the development of prototyping technology for the assessment of specific product functions. For this purpose, multiple sensory simulation was incorporated to mimic the product behavior and to create a realistic interaction to the prototypes [21,22]. Few studies have analyzed whether or not virtual prototypes are comparable with the real product on evaluation of basic product attributes. Understanding the difference, if there is any, between different prototypes with regard to the human perception of the product they represent is also valuable.

In this study, we conducted a holistic comparative study on the effectiveness of virtual prototypes versus the actual product they aim to represent for design evaluation purposes. At the first stage, experiments were carried out to understand the human subjects' perception of the physical and appearance features of the product, and the usability of operating the product's functions by using both media. The emotional responses to the performance of the evaluation tasks were also analyzed and compared on the basis of the experimental results. At the second stage, additional tests were conducted to demonstrate that incorporating an instant sensory feedback in visual virtual prototypes improves the subjects' emotional responses to the evaluation tasks. These findings may provide useful information to the refinement of virtual prototyping technology for design evaluation purposes and can also improve the user experience design of interactive functions for emerging virtual and augmented reality (VR/AR) applications.

2. Methodology

We conducted two-stage experiments to comprehensively investigate the users' responses to various design attributes of the virtual prototypes and the product they aimed to represent. The experimental procedure is shown in Figure 1. The first stage aims to identify potential differences between the effectiveness of design evaluation with a virtual prototype and actual product. A pretest was carried out to improve the original design of the experiments. Subjects evaluated both the virtual prototype and actual product from three aspects: Physical/appearance, usability, and emotional response by the measure of single-answer questions, procedural correctness, and an emotional scale, respectively. A statistical analysis of the experimental results helps identify significant differences existing in the two evaluation forms. Specific product functions showing the differences were also recognized from the analysis. The design of the questions (single answer with five options) was not justified compared to other alternatives. This is a limitation of the methodology used by this study.

The second stage is to compare the performance of a virtual prototype on those product functions integrated with different feedbacks. The comparison was focused on the aspects of usability and emotional response. Similar to the previous stage, they were measured by procedural correctness and the emotional scale, respectively. The experimental results were then analyzed to determine whether or not including the feedbacks improves the two measures.

The relative order of obtaining the SAM scale in the experimental procedure influences the effectiveness of evaluating emotional responses. As shown in Figure 1, subjects gave a score on the SAM scale right after they had determined a design attribute or operated a product function.

Under such circumstance, the time lag between experiencing an activity and subjectively assessing the experience is minimized. This justifies the use of the SAM scale.

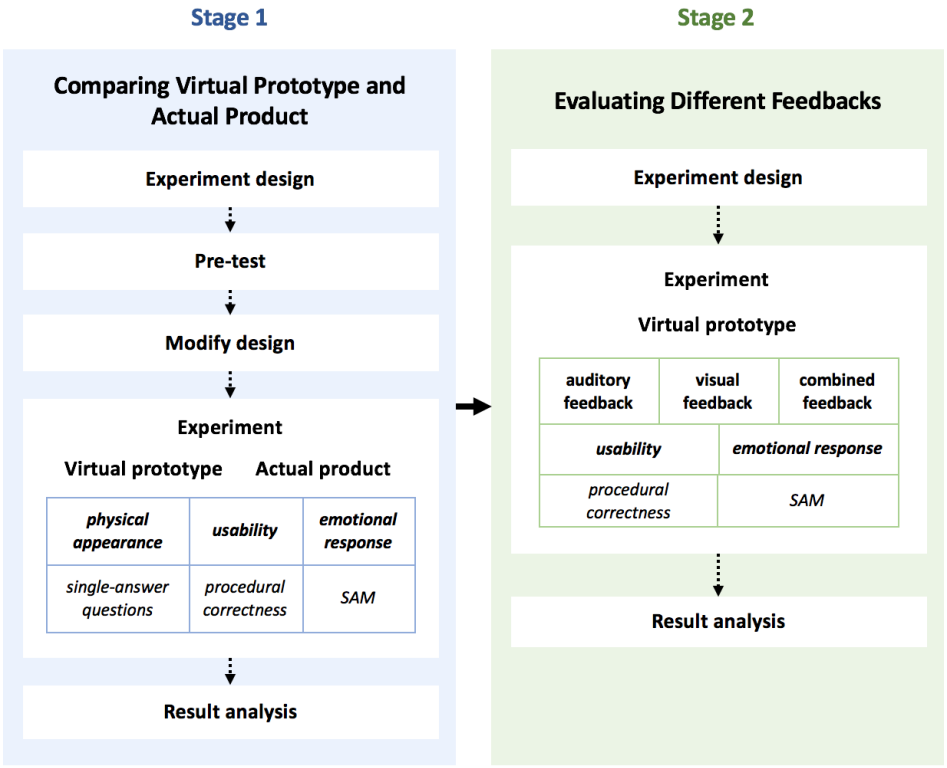


Figure 1. A two-stage experimental procedure proposed by this study.

In this study, we compared two product forms (real and virtual) from three aspects (functional, usability, emotional) using both quantitative (single-selection questions, procedure correctness) and qualitative measures (SAM), respectively. As shown in Figure 2, experiments were systematically carried out to understand the advantages of using an actual representative product over using visual virtual prototypes. The analysis of experimental data can reveal the potential means of improving the virtual prototypes in the evaluation of product design.

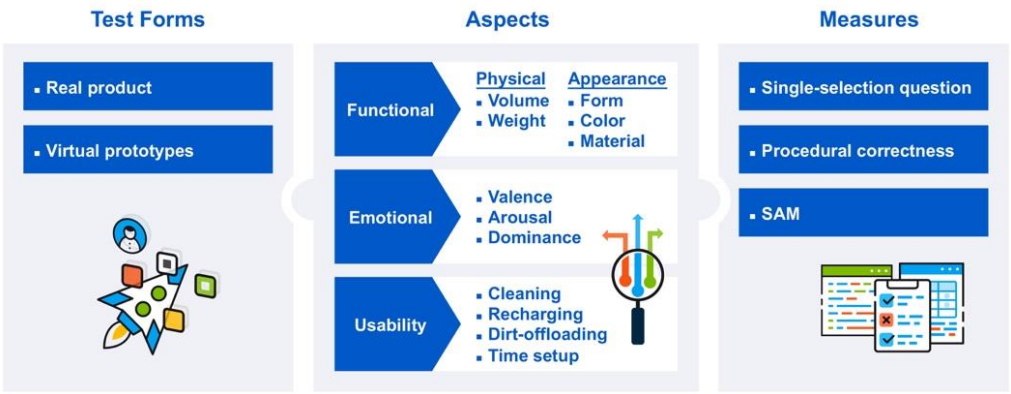


Figure 2. Systematic design of experiments.

The users' perception of the product's dimensions, shape, and weight is critical with regard to influencing the user's experience and purchase decision of the product [23]. It is important to understand whether or not product users can correctly recognize the product's physical attributes

such as volume and weight. Design evaluation is also highly related to the human's perception of the product shape, color, and material. In the experiments, subjects evaluated those physical and appearance attributes by answering a single selection question for each attribute (see Appendix A). They operated specific product functions with both the virtual prototype and actual product in the usability assessment. The operation of each function follows a well-defined procedure. The product usability was measured by the procedural correctness. In addition, we assessed the subjects' emotional responses during the attribute evaluation processes and the operation procedures mentioned above.

2.1. Aspects of Design Evaluation in the Experiment

2.1.1. Evaluation Attributes in Physical Aspect

A design evaluation method should allow people to easily estimate a product from various physical aspects. Each subject has to determine three quantitative values for the given prototypes considered in the experiments:

1. Unreferenced volume: Estimating a product's volume without a reference object.
2. Referenced volume: Estimating a product's volume with a reference object. A 150-mm ruler is used as a reference in the experiments.
3. Weight: Estimating a product's weight.

2.1.2. Evaluation Attributes in Appearance Aspect

Form, color, and material are three commonly known elements constituting the appearance of a product. These elements are prominent factors signifying the perceptual responses of a user to a product [23]. In this study, the test product was relatively simple with regard to changing the three elements.

1. Form: How are the form features or shapes perceived?
2. Color: How are the colors perceived?
3. Material: How are the materials perceived?

2.1.3. Evaluation Attributes in Usability Aspect

Usability is related to the performance of a specific task or product function. From this viewpoint, we devised the following user tasks to allow the participants to operate the prototype of an automatic cleaning device.

1. Cleaning: Activating the cleaning process.
2. Recharging: Sending the device back to the charging base and recharging it.
3. Time setup: Setting up the cleaning time.
4. Dirt offloading: Removing the collected dirt from within the container of the device.

2.1.4. Evaluation Attributes in Emotional Aspect

It would be interesting to recognize the emotional responses of a subject when conducting individual evaluation tasks. We employed the self-assessment manikin (SAM) scale, which was developed by Bradley and Liang [24]. This is a visually-based emotional scale developed according to the three emotional constructs proposed by Russell and Mehrabian [25]. For each construct, there exists five pictorial manikins, each of which is accompanied by a nine-level scale. As shown in Figure 3, the three constructs are:

1. Valence (negative-positive),
2. Arousal (passive-active), and
3. Dominance (dominated-dominant).

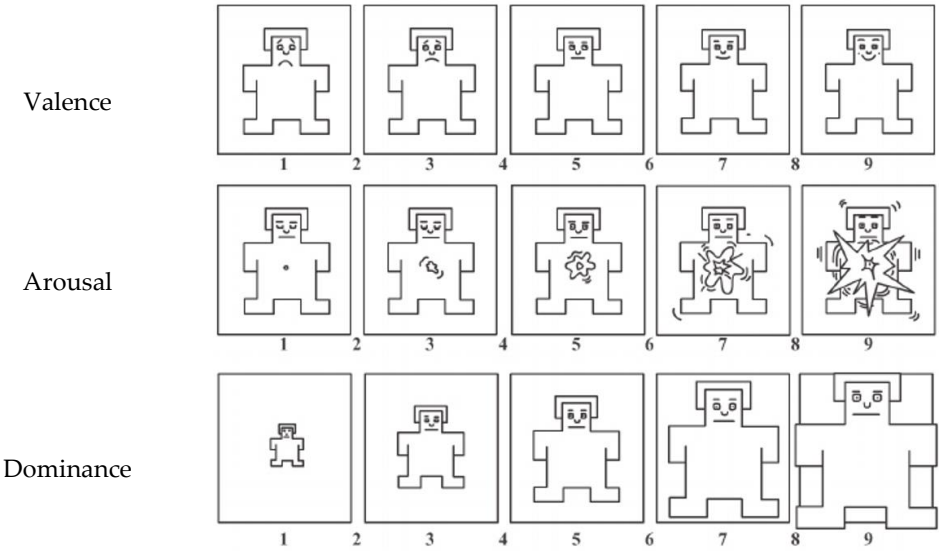


Figure 3. Self-assessment manikin (SAM) scale [24].

SAM has been used in several design-related studies to evaluate the users' emotional responses to products [26]. Its visual approach can facilitate cross-cultural studies of product evaluation without the need of dealing with different languages [27]. Moreover, SAM is freely available and straightforward to use.

2.2. Experiment Design

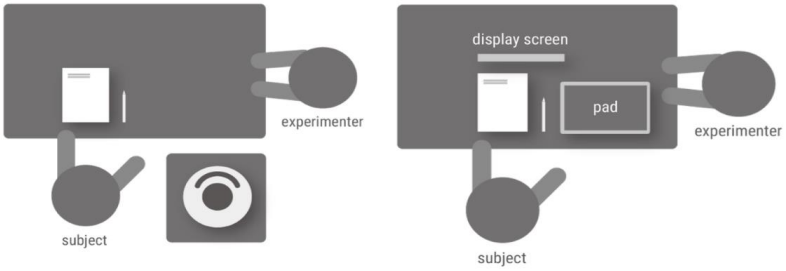
As shown in Figure 4, the representative product used in this study was a smart vacuum cleaner (Roomba5815, produced by iRobot). When activated, this smart device collects dirt from the floor by travelling around the house while avoiding obstacles. We selected this product because the participants were expected to have limited knowledge and experience of using this type of product. Therefore, the confounding effect of prior experience was reduced.

Figure 4. Smart vacuum cleaner as the test product.

We recruited a group of 50 different Taiwanese people in operating the test product and a virtual prototype, respectively. There were 100 participants involved at the first stage of the experiments with an equal female-to-male ratio. Those participants were aged from 20 to 29. Among them 82 people were college students and the others were engineers with one to four-year work experience. This group of people were not professional cleaners nor mainly responsible for house cleaning work in their households. They were less likely to have used the test product. The confounding effect of prior experience was thus reduced. However, they did have substantial experience in using electronic devices and were familiar with the icons and symbols commonly used on such devices. At the second

stage of the experiments, a group of 30 different people conducted the design evaluation with auditory, visual, and combined feedback, respectively. Those 90 participants have the same profile as the first stage. In summary, there were two groups of 50 people in the first test and three groups of 30 people in the second one.

The experiments were carried out in a quiet room with sufficient lighting and space for the participants to comfortably perform the experimental tasks. In an orientation session prior to the experiment, the experimenter explained the procedure and testing rules, and the participants were free to ask questions during the session. Talking was not allowed during the actual experiment. The experimental setup for the physical and virtual groups, respectively, will now be described. As shown in Figure 5a, the smart vacuum cleaner was placed in a low platform close to the ground. The participants inspected the device and performed the required tasks. They could freely lift the product with hands during the experiment for the weight estimation. Then, they responded to the questionnaire, including the SAM, by filling in the answer sheets that were placed on the desk.



(**a**) (**b**)

Figure 5. Experimental setting for (**a**) physical product and (**b**) virtual prototypes.

A highly realistic 3D model was developed as the visual virtual prototype of the smart device (Figure 6). We used the 3D rendering software Keyshot 5 to render and display the model. The browsing mode in the software allowed the participants to freely rotate the prototypes in 3D space. Several interactive functions were implemented by using Axure RP (https://www.axure.com/) to provide the user feedback that will be described in the next section. The software ran on a tablet device. When performing the usability tasks, the participants touched the pad's screen to operate the prototype's interface. The participants were then requested to respond according to the SAM scale.

(**a**) (**b**)

Figure 6. (**a**) Physical product vs. (**b**) virtual prototype.

2.3. Experimental Procedure

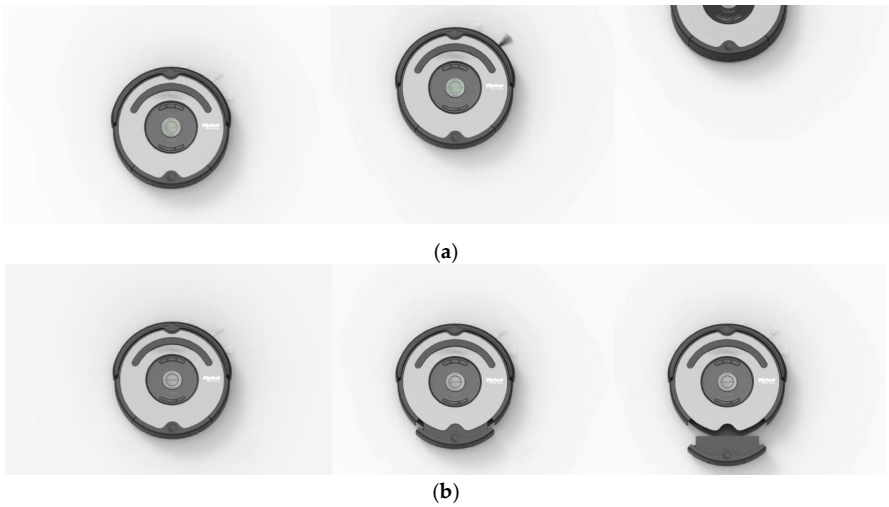
- First Stage: Comparison between the Real Product and Virtual Prototype

The first part of the experiments aimed to determine to which extent the visual virtual prototypes were different from the physical product. We asked the participants to estimate the physical dimensions and appearance of the two media and used them to carry out the usability tasks. They needed to fill out the SAM sheet after completing each task.

- Second Stage: Tests with Additional Feedback

At the second stage, the tasks with significant differences were selected as the user tasks for which additional feedback would be provided. They include auditory, visual, and combined feedback, and are described as follows:

1. Auditory Feedback: When pressing the power button, a beeping sound was reproduced. After the entire task had been completed, a three-note sound effect was reproduced.
2. Visual Feedback: When pressing the power button, an indication light was turned on. After the entire task had been completed, the device moved up to the top of the pad's screen and its brush started to rotate (Figure 7a). For the dirt-offloading task, the dirt container moved out of the device (Figure 7b).



(**a**)

(**b**)

Figure 7. Visual virtual prototype responses with various feedback to (**a**) cleaning and (**b**) dirt- offloading.

Combined Feedback: After the same operations had been carried out by the participants, both the auditory and visual feedback was activated during the execution of the tasks. Before conducting the actual experiments, we conducted a pilot study to test the experimental environment and the tasks assigned to the users. Four participants were recruited to attend the pilot study; that is, two participants for each group. The actual experiments were improved based on their suggestion. For example, the platform was adjusted to a lower position. Prior to the experiments, a verbal explanation was provided by the conductor, who also suggested to include a sample question and its answer in order to demonstrate the response process. The task instructions were re-written in a clear and concise manner.

In the experiments regarding the physical and appearance aspects, the participants were requested to fill in two questionnaires: The answer sheet for the evaluated product features and the SAM scale. In the usability tests, the accuracy with which the participants performed the tasks was determined directly from the test results. The SAM was also used to monitor the emotional responses. Instead of asking the participants to fill in the names of product attributes, we used single-choice questions. The main questions were listed with five optional answers. The participants selected the answer they believed to be the correct one.

3. Results and Discussion

3.1. Physical Aspects

- Unreferenced Volume

Accuracy: The physical group achieved a higher accuracy of 62% in comparison to the virtual group's accuracy of 53% with regard to volume estimation. However, the results obtained with the chi-square test revealed that this difference was not statistically significant.

Emotional responses: Only the valence result was statistically significant. The physical group was in a relatively more positive state.

- Referenced Volume

Accuracy: Both groups achieved a higher level of accuracy (physical: 72%, virtual: 64%), as shown in Table 1a,b. The accuracy of the virtual group increased to a level similar to that of the physical group without the reference. This may indicate that adding a reference object could enhance the prediction capability of the virtual prototypes. The chi-square test result revealed that the performance of both groups was significantly different (see Table 1c). The physical group tended to underestimate the volume, while the virtual group tended to overestimate it.

Table 1. (**a**) Estimation of referenced volume with the physical product; (**b**) estimation of referenced volume with the virtual prototype; (**c**) chi-square test for the referenced volume estimation.

(**a**)		
Option	**Number of Times**	**%**
2	13	26.0
3 (correct)	36	72.0
4	1	2.0
Total	50	100.0

(**b**)		
Option	**Number of Times**	**%**
2	6	12.0
3 (correct)	32	64.0
4	12	24.0
Total	50	100.0

(**c**)			
	Values	**DF**	**Sig.**
Chi-Square test	12.122 [a]	2	0.002

[a] 0 cells have count less than 5.

Emotional responses: The two groups had no significant differences.

- Weight

Accuracy: Both groups achieved a relatively low accuracy. The chi-square test result revealed that the performances were not significantly different, and the physical and virtual features did not seem to provide a good hint for weight estimation.

Emotional responses: Only the valence construct was significantly different between the two groups. Using the physical product seemed to invoke a more positive emotional response in comparison with using the visual virtual prototype.

3.2. Appearance Aspect

- Form

Accuracy: Both groups performed very well by achieving over 90% accuracy. The chi-square test result indicated no significant difference between the two groups.

Emotional responses: Only the valence construct was significantly different between the two groups. The physical group was in a more positive emotional state.

- Color

Accuracy: The physical group achieved higher accuracy (62%) in comparison to the virtual group (48%). This result may be natural because the real product should, by all means, have the "right" color. The chi-square test result revealed that the difference was not statistically significant.

Emotional responses: The two groups reacted differently in terms of the valence construct and arousal construct. The physical group was in a more positive state when evaluating the colors of the physical product, while the virtual group was in a more aroused state when evaluating the colors of the virtual prototype.

- Material

Accuracy: The physical group performed rather well, with an accuracy of 96%, while the virtual group achieved an accuracy of 76% (see Table 2a,b). The chi-square test result shows that this difference is statistically significant (Table 2c).

Emotional responses: As shown in Table 2d, the two groups had a difference with regard to the valence construct. The physical group was in a more positive state.

Table 2. (**a**) Estimation of material with the physical product. (**b**) Estimation of material with the virtual prototype. (**c**) Chi-square test for estimation of material. (**d**) ANOVA for estimation of material.

(**a**)

Option	Number of Times	%
1	2	4.0
2 (correct)	48	96.0
total	50	100.0

(**b**)

Option	Number of Times	%
1	12	24.0
2 (correct)	38	76.0
total	50	100.0

(**c**)

	Values	df	Sig.
Chi-Square test	8.306 [a]	1	0.004

(**d**)

		SS	DF	MS	F	P
Valence	Between Groups	19.360	1	19.360	10.218	0.002
	Within Groups	185.680	98	1.895		
	Total	205.040	99			
Arousal	Between Groups	7.840	1	7.840	2.366	0.127
	Within Groups	324.800	98	3.314		
	Total	332.640	99			
Dominance	Between Groups	9.610	1	9.610	3.282	0.073
	Within Groups	286.980	98	2.928		
	Total	296.590	99			

[a] 0 cells have count less than 5.

3.3. Usability Aspects

- Cleaning

Accuracy: The correct-hit rates of the physical and virtual groups were 80% and 86%, respectively (see Table 3a,b). Interestingly, the virtual group performed better in this task.

Table 3. (**a**) Usability assessment of the cleaning task with the physical product. (**b**) Usability assessment of the cleaning task with the virtual prototype. (**c**) ANOVA for usability assessment of the cleaning task.

(**a**)

Option	Number of Times	%
correct	40	80.0
incorrect	10	20.0
total	50	100.0

(**b**)

Option	Number of Times	%
correct	43	86.0
incorrect	7	14.0
total	50	100.0

(**c**)

		SS	DF	MS	F	P
Correct-hit	Between Groups	0.040	1	0.040	0.135	0.714
	Within Groups	29.120	98	0.297		
	Total	29.160	99			
Valence	Between Groups	100.000	1	100.000	28.229	0.000
	Within Groups	347.160	98	3.542		
	Total	447.160	99			
Arousal	Between Groups	60.840	1	60.840	19.644	0.000
	Within Groups	303.520	98	3.097		
	Total	364.360	99			
Dominance	Between Groups	0.010	1	0.010	0.002	0.961
	Within Groups	416.900	98	4.254		
	Total	416.910	99			

Emotional responses: The two groups were significantly different in the valence and arousal constructs (Table 3c). The physical group was in a more positive and aroused state, while the virtual group was in a more negative and passive state.

- Recharging

Accuracy: Both groups achieved high correct-hit rates (physical: 98%, virtual: 96%) as shown in Table 4a,b, respectively. Recharging appeared to be a well-designed operating feature and easy for the participants to execute.

Emotional responses: The physical group was in a more positive and aroused state. The virtual group was in a more negative and passive state. The two groups had significant differences with regard to the valence and arousal constructs (see Table 4c).

Table 4. (**a**) Usability assessment of the recharging task with the physical product. (**b**) Usability assessment of the recharging task with the virtual prototype. (**c**) ANOVA for usability assessment of the recharging task.

(**a**)

Option	Number of Times	%
correct	49	98.0
incorrect	1	2.0
total	50	100.0

(**b**)

Option	Number of Times	%
correct	48	96.0
incorrect	2	4.0
total	50	100.0

(**c**)

	SS	DF	MS	F	P
Correct-hit	Between Groups	0.010	1	0.010	0.338	0.562
	Within Groups	2.900	98	0.030		
	Total	2.910	99			
Valence	Between Groups	23.040	1	23.040	10.982	0.001
	Within Groups	205.600	98	2.098		
	Total	228.640	99			
Arousal	Between Groups	19.360	1	19.360	6.244	0.014
	Within Groups	303.880	98	3.101		
	Total	323.240	99			
Dominance	Between Groups	0.810	1	0.810	0.223	0.638
	Within Groups	355.940	98	3.632		
	Total	356.750	99			

- Time setup

Accuracy: The participants were requested to set the time to Wednesday, 4:05 AM. The correct-hit rate of the physical and virtual groups was 96% and 86%, respectively.

Emotional responses: The two groups had no significant differences in their emotional states.

- Dirt-offloading

Accuracy: The virtual group achieved a higher correct-hit rate of 62% than the physical group's rate of 58% (see Table 5a,b). This could have been caused by the mechanism of the physical product requiring more manual work. However, for the virtual prototypes, a simple clicking action could accomplish the same task.

Emotional responses: The two groups reacted differently with regard to the valence and arousal constructs (see Table 5c). The physical group was in a more positive and aroused state, while the virtual group was in a more negative and passive state.

Table 5. (**a**) Usability assessment of the dirt-offloading task with the physical product. (**b**) Usability assessment of the dirt-offloading task with the virtual prototype. (**c**) ANOVA for usability assessment of the dirt-offloading task.

(**a**)			
	Option	**Number of Times**	**%**
	correct	29	58.0
	incorrect	21	42.0
	total	50	100.0
(**b**)			
	Option	**Number of Times**	**%**
	correct	31	62.0
	incorrect	19	38.0
	total	50	100.0

(**c**)

		SS	**DF**	**MS**	**F**	**P**
Correct-hit	Between Groups	0.040	1	0.040	0.059	0.809
	Within Groups	66.600	98	0.680		
	Total	66.640	99			
Valence	Between Groups	31.360	1	31.360	10.912	0.001
	Within Groups	281.640	98	2.874		
	Total	313.000	99			
Arousal	Between Groups	10.890	1	10.890	2.855	0.074
	Within Groups	373.860	98	3.815		
	Total	384.750	99			
Dominance	Between Groups	0.810	1	0.810	0.140	0.709
	Within Groups	565.380	98	5.769		
	Total	566.190	99			

Table 6 summarizes all the evaluation tasks in which the actual product and visual virtual prototypes exhibited a statistically significant difference. The estimation results for the product volume, color, and material of the virtual prototype produced lower accuracy results in comparison with the estimation results for the same aspects of the actual product. However, the latter failed to give a satisfactory degree of correctness (Sections 3.1 and 3.2). One possible reason is that, regardless of whether the product was physical or virtual, to estimate the product dimensions and to recognize its color and material, practice and/or specialized training is required, which the participants did not have. Additionally, the actual product caused a more positive and aroused emotional state, when performing the evaluation tasks. The virtual product produced a more negative and passive state. Although the visual quality of the virtual prototype was highly realistic (Figure 4), the participants might have expected additional sensory stimulus that matched their previous experience of interacting with actual products.

Table 6. Experimental results with significant differences.

Correctness	**Emotional-Valence**	**Emotional-Arousal**
• Referenced volume • Color • Material	• Unreferenced volume • Weight • Form • Color • Material • Cleaning • Recharging • Dirt-offloading	• Color • Cleaning • Recharging • Dirt-offloading

3.4. Providing Feedback

The results from the evaluation of usability revealed that the visual virtual prototypes invoked more negative and passive emotional responses, particularly with regard to the operations related to the cleaning and dirt-offloading tasks. It was advantageous to investigate whether adding an instant sensory feedback to these tasks could enhance the user experience. Thus, we incorporated auditory, visual, and combined feedback during the operations. The same usability tests (Section 3.3) were conducted again. The objective was to determine whether various types of feedback could exert a different influence on the emotional responses and task performance. Not all feedback had a significant influence on the emotional responses or task performance with regard to the cleaning task. Table 7 indicates a significantly different arousal state for the three feedback groups with regard to their dirt-offloading task responses. The combined feedback caused the highest level of arousal.

Table 7. ANOVA for the dirt-offloading task with different feedbacks.

		SS	DF	MS	F	P
Correct-hit	Between Groups	0.067	1	0.067	0.051	0.821
	Within Groups	75.267	58	1.298		
	Total	75.333	59			
Valence	Between Groups	5.400	1	5.400	2.692	0.106
	Within Groups	116.333	58	2.006		
	Total	121.733	59			
Arousal	Between Groups	14.017	1	14.017	5.140	0.027
	Within Groups	158.167	58	2.727		
	Total	172.183	59			
Dominance	Between Groups	1.067	1	1.067	0.250	0.619
	Within Groups	247.667	58	4.270		
	Total	248.733	59			

SS: Sum of squares; DF: Degrees of freedom; MS: Mean square.

Subsequently, we gathered the data obtained by the first experiment for the same two tasks and analyzed the responses of the three feedback groups by using hypothesis testing. To facilitate the analysis, we encoded the groups with the following values: Physical group: +1; visual virtual group: −1; auditory feedback: −2; visual feedback: −3; combined feedback: −4. When conducting the following multiple comparisons, we used the combined group (−4) as the base reference.

- Cleaning

Accuracy: There was no significant difference in the performance of the five groups.

Emotional responses (valence): According to Table 8a, there was a significant difference among the five groups with regard to the valence construct. From the multiple comparisons listed in Table 8(b), we could observe that differences existed between the combined feedback group and the visual virtual group without any feedback (−1). The addition of multiple sensory feedback improved the valence construct when the two tasks were carried out by using the virtual prototype. The combined group had the most positive emotional response.

Emotional responses (arousal): Table 9a indicates that there was a significant difference in the arousal construct among the groups. Table 9b shows that differences existed between the reference group and the auditory feedback group. The physical group had the highest level of arousal, while the combined feedback group had the second highest level of arousal, which was still higher than that of the visual virtual group. This indicates that the addition of combined feedback increased the arousal state to a level closer to that of the physical group.

Table 8. (**a**) ANOVA for the valence construct of the cleaning task with different feedbacks. (**b**) Multiple comparisons of the valence construct for integrated feedback.

(**a**)					
Source	**DF**	**Adj SS**	**Adj MS**	**F**	**P**
Regression	10	161.756	16.176	5.93	0.000
Virtual versus Physical	4	121.577	30.394	11.15	0.000
Error	179	488.054	2.727		
Pure Error	116	295.566	2.548		
Total	189	649.811			
(**b**)					
Term	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	1.81	1.08	1.67	0.096	
Virtual versus Physical					
−3	0.073	0.475	0.15	0.878	2.09
−2	0.325	0.446	0.73	0.467	1.84
−1	2.822	0.692	4.08	0.000	6.47
1	0.903	0.757	1.19	0.234	7.73

SS: Sum of squares; DF: Degrees of freedom; MS: Mean square. VIF: Variance inflation factor.

Table 9. (**a**) ANOVA for the arousal construct of the cleaning task with different feedbacks. (**b**) Multiple comparisons of the arousal construct for integrated feedback.

(**a**)					
Source	**DF**	**Adj SS**	**Adj MS**	**F**	**P**
Regression	10	87.470	8.747	2.65	0.005
Virtual versus Physical	4	71.840	17.960	5.43	0.000
Error	179	591.772	3.306		
Pure Error	116	382.664	3.299		
Total	189	679.242			
(**b**)					
Term	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	4.46	1.19	3.74	0.000	
Virtual versus Physical					
−3	0.724	0.523	1.38	0.168	2.09
−2	0.948	0.491	1.93	0.055	1.84
−1	0.732	0.762	0.96	0.338	6.47
1	−0.881	0.833	−1.06	0.291	7.73

SS: Sum of squares; DF: Degrees of freedom; MS: Mean square. VIF: Variance inflation factor.

Emotional responses (dominance): There was no significant difference in the groups with regard to the dominance construct.

- Dirt-offloading

Accuracy: In terms of accuracy, there was no significant difference in the performances of the five groups.

Emotional responses (valence): Table 10a indicates that there were significant differences in the responses of the five groups with regard to the valence construct. According to the multiple comparisons shown in Table 10), the responses of the combined feedback group were different than those of the visual virtual (−1) and auditory feedback (−2) groups. The combined feedback group produced a valence level that was higher than that of the physical group. All the feedback groups were in a more positive state in comparison to the visual virtual group.

Table 10. (**a**) ANOVA for the valence construct of the dirt-offloading task with different feedbacks. (**b**) Multiple comparisons of the valence construct for integrated feedback.

(**a**)					
Source	**DF**	**Adj SS**	**Adj MS**	**F**	**P**
Regression	10	103.240	10.3240	0.445	0.000
Virtual versus Physical	4	53.566	13.3915	5.57	0.000
Error	179	430.702	2.4062		
Pure Error	116	249.247	2.1487		
Total	189	533.942			
(**b**)					
Term	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	3.79	1.02	3.72	0.000	
Virtual versus Physical					
−3	0.693	0.446	1.55	0.122	2.09
−2	0.949	0.419	2.27	0.025	1.84
−1	1.683	0.650	2.59	0.010	6.47
1	0.509	0.711	0.72	0.475	7.73

SS: Sum of squares; DF: Degrees of freedom; MS: Mean square. VIF: Variance inflation factor.

Emotional responses (arousal): From Table 11a, it can be seen that there was a significant difference in the responses of the five groups with regard to the arousal construct. Specifically, the difference existed between the combined feedback group and the visual feedback group (Table 11b). The combined group had the highest level of arousal.

Table 11. (**a**) ANOVA for the arousal construct of the dirt-offloading task with different feedbacks. (**b**) Comparison of the arousal construct for integrated feedback.

(**a**)					
Source	**DF**	**Adj SS**	**Adj MS**	**F**	**P**
Regression	10	81.603	8.1603	2.43	0.010
Virtual versus Physical	4	30.030	7.5075	2.23	0.067
Error	179	602.166	3.3641		
Pure Error	116	381.915	3.2924		
Total	189	683.768			
(**b**)					
Term	Coef	SE Coef	T	P	VIF
Constant	3.77	1.20	3.13	0.002	
Virtual versus Physical					
−3	0.949	0.527	1.80	0.074	2.09
−2	0.629	0.495	1.27	0.205	1.84
−1	1.264	0.769	1.64	0.102	6.47
1	0.554	0.840	0.66	0.511	7.73

SS: Sum of squares; DF: Degrees of freedom; MS: Mean square. VIF: Variance inflation factor.

Emotional responses (dominance): A significant difference was not observed in the responses of the five groups with regard to the dominance construct.

In summary, the addition of combined feedback (visual and auditory) produced a statistically significant improvement with regard to the valence and arousal constructs, when performing the cleaning and dirt-offloading tasks with the virtual prototypes. The emotional states created in this manner could match those induced by using the actual product in the test results. In the experiments,

an interesting observation was that most participants appeared to be positively surprised when they received feedback. Therefore, sensory feedback may have helped the virtual prototypes mimic the behavior of the actual product to a larger degree. A similar observation was reported in the previous study of product use experience with an integrated with sensory feedback [28].

To test the reliability of the questionnaires, we analyzed the questionnaires collected from the physical and virtual groups with Cronbach's Alpha test. The results revealed that the alpha value of the physical group's questionnaires was 0.709, while that of the virtual group was 0.759. The value of all the questionnaires as a whole was 0.718. These numbers indicate the high reliability of the questionnaires and imply good consistency for the responses collected from different participants.

4. Conclusions and Future Work

The effectiveness of using virtual prototypes in design evaluation remains an active research topic. In this study, we conducted a comprehensive and systematic investigation to understand how effective visual virtual prototypes work in design evaluation in comparison with using a smart vacuum cleaner as a representative product. A series of experiments were devised to understand how well each product form (real and virtual) performs on design evaluation from three aspects (functional, usability, emotional) using both objective (single-selection questions, procedure correctness) and subjective measures (SAM), respectively. The motivation was to derive useful findings from the statistical analysis of the experimental data. They may benefit the development of product prototyping methods from the perspective of user-centric design evaluation. First, physical means still have some advantage in estimating the product volume, color, and material, in comparison to virtual means. This conclusion confirms the previous results [7,16]. Secondly, regardless of physical or virtual, to estimate the product dimensions and to recognize its color and material, practice and/or specialized training is required, which the participants did not have. The experimental result that both groups did not perform well in estimating the weight of the product may seem counter intuitive. We speculate that people have difficulty estimating a quantitative measure for weight. This issue was rarely mentioned by the related studies, though. In the usability tests, the visual virtual prototypes worked almost as well as the actual product in terms of using the product functions correctly. However, the two groups had different emotional responses when carrying out the same tasks. The visual virtual group without any sensory feedback generally tended to exhibit a more negative and passive emotional state. By adding user feedback, such as auditory and motion cues, the emotional responses towards the virtual prototypes changed to a level similar or higher to that of the physical product. The participants appeared to be positively surprised when they received feedback, which might have assisted the virtual prototypes in closely mimicking the behavior of the actual product. The past studies [26,28] suggested a similar effect of multimodal feedback in the product affective design.

Those findings may help improve the practicality of virtual prototyping in product development by adding useful features. They can also contribute to the design of interactive functions that will be suitable to emerging VR/AR applications, where user experience may be enhanced by integrating additional sensory feedback. To understand the effectiveness of product evaluation in VR /AR would be an interesting research topic. There are potential problems to be overcome in the research, though. The immersive experience created by a VR/AR environment may be too intense for subjects to focus on assessment measures related to design evaluation. The mental and physical workloads induced by the current devices can also bias the experimental result. Future work may also include a qualitative analysis to understand why users have certain emotional responses. Physiological measurements, such as eye tracking and the heart rate, can also be integrated to provide a more objective means of determining the users' mental state during the evaluation. Emotion analysis using the facial expression recognition software may also help characterize their emotional responses with objective measures during the experiment.

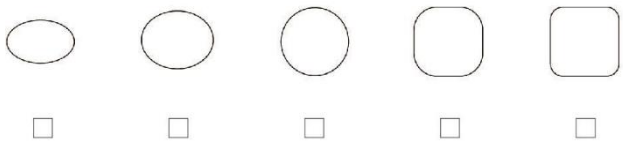
Author Contributions: Conceptualization, C.-H.C.; Data curation, E.-T.K.; Investigation, C.-H.C. and E.-T.K.; Methodology, C.-H.C.; Software, E.-T.K.; Validation, E.-T.K.; Writing – original draft, C.-H.C.; Writing—review & editing, C.-H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

- VOLUME Please carefully observe the product and select the one closest to your answer.
  - ○ 20 × 20 × 5 cm
  - ○ 25 × 25 × 7.5 cm
  - ○ 35 × 35 × 10 cm
  - ○ 45 × 4 × 12.5 cm
  - ○ 50 × 50 × 15 cm
- WEIGHT Please carefully observe the product and select the one closest to your answer.
  - ○ 2 kg
  - ○ 4 kg
  - ○ 8 kg
  - ○ 15 kg
  - ○ 30 kg
- SHAPE Please carefully observe the product and select the one closest to your answer.



- COLOR Please carefully observe the product and select the one closest to your answer.

- MATERIAL Please carefully observe the product and select the one closest to your answer.
  - ○ metal
  - ○ plastic
  - ○ wood
  - ○ glass
  - ○ leather

References

1. Arrighi, P.A.; Mougenot, C. Towards user empowerment in product design: A mixed reality tool for interactive virtual prototyping. *J. Intell. Manuf.* **2019**, *30*, 743–754. [CrossRef]
2. Lo, C.H.; Chu, C.H.; Huang, S.H. Evaluating the effect of interactions between appearance-related product designs and facial characteristics on social affectivity. *Int. J. Ind. Ergon.* **2019**, *45*, 35–47. [CrossRef]
3. Camburn, B.; Viswanathan, V.; Linsey, J.; Anderson, D.; Jensen, D.; Crawford, R.; Wood, K. Design prototyping methods: State of the art in strategies, techniques, and guidelines. *Des. Sci.* **2017**, *3*, e13. [CrossRef]

4. Jordan, P.W. *Designing Pleasurable Products: An Introduction to the New Human Factors*; CRC Press: Boca Raton, FL, USA, 2003.
5. Söderman, M. Virtual reality in product evaluations with potential customers: An exploratory study comparing virtual reality with conventional product representations. *J. Eng. Des.* **2005**, *16*, 311–328. [CrossRef]
6. Ibrahim, R.; Rahimian, F.P. Comparison of CAD and manual sketching tools for teaching architectural design. *Autom. Constr.* **2010**, *19*, 978–987. [CrossRef]
7. Gibson, I.; Gao, Z.; Campbell, I. A comparative study of virtual prototyping and physical prototyping. *Int. J. Manuf. Technol. Manag.* **2004**, *6*, 503–522. [CrossRef]
8. Pontonnier, C.; Dumont, G.; Samani, A.; Madeleine, P.; Badawi, M. Designing and evaluating a workstation in real and virtual environment: Toward virtual reality based ergonomic design sessions. *J. Multimodal User Interfaces* **2014**, *8*, 199–208. [CrossRef]
9. Kim, C.; Lee, C.; Lehto, M.R.; Yun, M.H. Affective evaluation of user impressions using virtual product prototyping. *Hum. Factor Ergon. Man.* **2011**, *21*, 1–13. [CrossRef]
10. Aromaa, S.; Väänänen, K. Suitability of virtual prototypes to support human factors/ergonomics evaluation during the design. *Appl. Ergon.* **2016**, *56*, 11–18. [CrossRef]
11. Ferrise, F.; Graziosi, S.; Bordegoni, M. Prototyping strategies for multisensory product experience engineering. *J. Intell. Manuf.* **2017**, *28*, 1695–1707. [CrossRef]
12. Faust, F.G.; Catecati, T.; de Souza Sierra, I.; Araujo, F.S.; Ramírez, A.R.G.; Nickel, E.M.; Ferreira, M.G.G. Mixed prototypes for the evaluation of usability and user experience: Simulating an interactive electronic device. *Virtual Real.* **2019**, *23*, 197–211. [CrossRef]
13. Reid, T.N.; MacDonald, E.F.; Du, P. Impact of product design representation on customer judgment. *J. Mech. Des.* **2013**, *135*, 091008. [CrossRef]
14. Chun, S.; Nam, K. User-centred design approaches for planning inpatient room of geriatric long-yerm care hospitals: Design factors with practical suggestions. *Des. J.* **2019**, *22*, 413.
15. Edler, D.; Keil, J.; Wiedenlübbert, T.; Sossna, M.; Kühne, O.; Dickmann, F. Immersive VR experience of redeveloped post-industrial sites: The example of Zeche Holland in Bochum-Wattenscheid. *J. Cartogr. Geogr. Inf.* **2019**, *69*, 267–284. [CrossRef]
16. Bligård, L.O.; Berlin, C.; Österman, C. The power of the dollhouse: Comparing the use of full-scale, 1:16-scale and virtual 3D-models for user evaluation of workstation design. *Int. J. Ind. Ergon.* **2018**, *68*, 344–354. [CrossRef]
17. Voit, A.; Mayer, S.; Schwind, V.; Henze, N. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019.
18. Kuliga, S.F.; Thrash, T.; Dalton, R.C.; Hölscher, C. Virtual reality as an empirical research tool—Exploring user experience in a real building and a corresponding virtual model, Comput. Environ. *Urban Syst.* **2015**, *54*, 363–375. [CrossRef]
19. Higuera-Trujillo, J.L.; Maldonado, J.L.T.; Millán, C.L. Psychological and physiological human responses to simulated and real environments: A comparison between Photographs, 360 Panoramas, and Virtual Reality. *Appl. Ergon.* **2017**, *65*, 398–409. [CrossRef]
20. Felnhofer, A.; Kothgassner, O.D.; Schmidt, M.; Heinzle, A.K.; Beutl, L.; Hlavacs, H.; Kryspin-Exner, I. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *Int. J. Hum. Comput. Stud.* **2015**, *82*, 48–56. [CrossRef]
21. Ha, S.; Kim, L.; Park, S.; Jun, C.S.; Rho, H. Virtual prototyping enhanced by a haptic interface. *CIRP Ann.* **2009**, *5*, 135–138. [CrossRef]
22. Rhiu, I.; Bahn, S.; Nam, C.S.; Yun, M.H. Affective experience of physical user interfaces: Similarities and differences among control types. *Hum. Factor Ergon. Man.* **2018**, *28*, 56–68. [CrossRef]
23. Hsiao, K.A.; Chen, L.L. Fundamental dimensions of affective responses to product shapes. *Int. J. Ind. Ergon.* **2006**, *36*, 553–564. [CrossRef]
24. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
25. Russell, J.A.; Mehrabian, A. Evidence for a three-factor theory of emotions. *J. Res. Pers.* **1977**, *11*, 273–294. [CrossRef]

26. Bertheaux, C.; Toscano, R.; Fortunier, R.; Borg, C. Integration of perception and emotions in a new sensory design process. *Int. J. Des. Eng.* **2018**, *8*, 73–97. [CrossRef]
27. Sonderegger, A.; Sauer, J. The influence of socio-cultural background and product value in usability testing. *Appl. Ergon.* **2013**, *44*, 341–349. [CrossRef]
28. Razza, B.M.; Paschoarelli, L.C.; Santos, H.M.; Andrade, L.O. The multisensory experience: A case study with five different products. *Adv. Ergon. Design Usability Spec. Popul.* **2014**.