

Article

Driving across Markets: An Analysis of a Human–Machine Interface in Different International Contexts

Denise Sogemeier ^{1,2,*} , Yannick Forster ¹ , Frederik Naujoks ¹ , Josef F. Krems ²  and Andreas Keinath ¹ 

¹ BMW Group, 80937 Munich, Germany; yannick.forster@bmw.de (Y.F.); frederik.naujoks@bmw.de (F.N.); andreas.keinath@bmw.de (A.K.)

² Department of Psychology, Chemnitz University of Technology, 09111 Chemnitz, Germany; josef.krems@psychologie.tu-chemnitz.de

* Correspondence: denise.sogemeier@bmw.de; Tel.: +49-151-601-25986

Abstract: The design of automotive human–machine interfaces (HMIs) for global consumers’ needs to cater to a broad spectrum of drivers. This paper comprises benchmark studies and explores how users from international markets—Germany, China, and the United States—engage with the same automotive HMI. In real driving scenarios, $N = 301$ participants (premium vehicle owners) completed several tasks using different interaction modalities. The multi-method approach included both self-report measures to assess preference and satisfaction through well-established questionnaires and observational measures, namely experimenter ratings, to capture interaction performance. We observed a trend towards lower preference ratings in the Chinese sample. Further, interaction performance differed across the user groups, with self-reported preference not consistently aligning with observed performance. This dissociation accentuates the importance of integrating both measures in user studies. By employing benchmark data, we provide insights into varied market-based perspectives on automotive HMIs. The findings highlight the necessity for a nuanced approach to HMI design that considers diverse user preferences and interaction patterns.

Keywords: benchmarking; human–machine interaction; human–machine interface; usability; user experience



Citation: Sogemeier, D.; Forster, Y.; Naujoks, F.; Krems, J.F.; Keinath, A. Driving across Markets: An Analysis of a Human–Machine Interface in Different International Contexts. *Information* **2024**, *15*, 349. <https://doi.org/10.3390/info15060349>

Academic Editor: Gerrit Meixner

Received: 22 April 2024

Revised: 24 May 2024

Accepted: 4 June 2024

Published: 12 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Watson [1] emphasized quality, technology, and cost as crucial for surpassing global competition. To this day, the urgency to outperform competitors has further intensified. Benchmarking is a common method that allows companies to evaluate their products by comparing them to those of their direct competitors. While conventional usability studies aim at improving a product and eliminating errors in the development cycle [2], benchmarking is defined as a process of identifying, understanding, and adapting outstanding practices from organizations anywhere in the world. Benchmark studies pursue the overachieving goal of improving performance [3] and ensuring customer satisfaction [4]. The findings provide important impulses for product requirements in a product’s development cycle. Hence, benchmarking can be recognized “as a catalyst for improvement and innovation” [5] (p. 258). Furthermore, by benchmarking, researchers can identify best practices and industry standards that contribute to optimal human–system interaction. This knowledge helps in advancing the field of human factors by providing insights into effective design principles, ergonomic considerations, and user-centered approaches.

1.1. Evaluation and Development of Human–Machine Interfaces

Benchmarking can be conducted in a number of contexts, including the automotive industry. In the automotive context, it is of particular importance to evaluate human–machine interfaces (HMIs) since their design is critical in several aspects of safety, efficiency, and branding [6]. Essential components of a vehicle’s HMI are the in-vehicle information

systems (IVISs), which include functions such as navigation, media, radio, and communication, among others. These systems should be used safely while driving since interaction with them is only a secondary or tertiary task [7]. Sufficient mental resources should remain for the primary driving task [8]. Apart from safety considerations, IVISs should be easy to use and understand. Additionally, as automated driving systems (ADSs) become increasingly accessible to a large consumer population, it becomes imperative to assess the appropriateness of HMIs and to evaluate them accordingly [9]. Those systems are evaluated during the development cycle [10] and can also be compared to those of competitors in in-market evaluations. The latter refers to benchmarking. When comparing systems, different constructs can be evaluated to assess the user's attitude towards the system. Commonly investigated constructs are acceptance, trust, usability, or user experience (UX) [11,12]. Usability is defined as “[...] the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [10]. UX expands the concept of usability by adding hedonic qualities such as joy and, thereby, going beyond mere pragmatic aspects [13]. Both concepts can be evaluated using standardized instruments such as rating scales and self-report questionnaires and can also be applied to evaluate ADSs [14]. Furthermore, observational measures should be included in user studies to derive a holistic picture of the in-vehicle HMI [15].

1.2. Benchmarking in Different Markets

As the automotive industry operates on a global level, benchmarking in an international context is essential. With respect to automotive HMIs, including ADSs, the design and development focus predominantly on the needs and preferences of drivers from Western markets [16]. Markets that differ from Western societies in, e.g., culture, language, driving environment, and behavior, may react differently to HMI solutions regarding their comprehension of and attitude towards the system. Designing a universal user interface (UI) that achieves comparable levels of user satisfaction is an immense challenge for original equipment manufacturers (OEMs) [17]. Regulation and legislation vary among countries, requiring market-specific solutions. The fact that market requirements differ also calls for specific solutions rather than a “one size fits all” HMI [6]. Differences in UI evaluation between countries have been acknowledged in previous research [17,18], showing that behavioral tendencies are associated with culture-influenced usability and design preferences [19]. Wang and colleagues [20], for example, found differences between Swedish and Chinese drivers regarding information requirements for advanced driving assistance systems. In complex traffic situations, the two cultural groups required different types of information. Khan and Williams [21] showed differences in expectations for HMI systems between Indian and British drivers: while usability was seen as a universal requirement to provide a satisfying product, Indian users held stronger opinions on functionality requirements such as user help. Similar results provided by Young and colleagues [22] showed differences between Australian and Chinese drivers in the comprehension of IVIS functions. Braun and colleagues [23] suggested considering cultural differences when designing effective automotive UIs, as these interfaces are perceived differently by German and Chinese users.

1.3. Research Question

Previous studies identified differences between different international user groups with prototypes in driving simulators using questionnaires. However, there is a dearth of research on how users from different markets perceive the same automotive UI integrated into a series production vehicle in a real driving context. As data derived from benchmark studies hold immense value, this paper attempts to process these datasets in a way that makes them not only accessible but also useful for research purposes. Over a span of three years, we collected diverse samples from three international markets, resulting in a substantial sample size. Subsequently, we meticulously processed and analyzed these datasets to extract greater utility and insights from the information they encapsulate. Does a global

user exist or are market-specific UI solutions in the automotive context promising? In order to gain a deeper understanding of this question and to identify potential avenues for further inquiry, we adopted an exploratory approach to target human–machine interaction with a focus on usability and UX. Besides self-report measures, the multi-method approach also included an observational measure as an indicator of interaction performance. The aim of this paper is to utilize benchmark data in order to provide preliminary insights into how user groups with distinct international backgrounds differ across a range of constructs. Based on the literature presented in Section 1.2, we assume that there are differences between user groups from three markets regarding self-report measures, i.e., satisfaction, hedonic qualities, overall UI evaluation, and interaction performance. We further propose that these differences may be attributed to different cultural standards, habits, and expectations prevalent in the different markets. Thereby, this paper aims to spark discussion about the underlying factors contributing to differences between international markets, addressing broader questions related to the root cause of such differences. By prioritizing the understanding of user preferences and interaction patterns across international markets, automotive manufacturers, designers, and researchers can develop more user-centric, globally relevant, and successful products and services. This deep user understanding can inspire new ideas and innovations that may not have been apparent from a single-market perspective. Cross-cultural insights can spark novel solutions to meet the evolving needs of automotive customers worldwide.

2. Materials and Methods

To gain further insights into whether real driving scenarios engender differences in the HMI evaluation, we used a multi-study research approach. An overview of the multi-study approach can be found in Table 1. Over the course of four years, we conducted a series of studies in three distinct markets, which were subsequently consolidated into three comprehensive studies labeled study 1, study 2, and study 3. We pursued similar procedures in each study and implemented a pre-defined set of use cases (UCs) and comparable measures. In study 2 and study 3, the vehicles tested were identical, whereas the vehicle in study 3 was equipped with an updated interface version. Consequently, comparisons were not performed across the three studies but within each study between the user groups. The primary difference in the updated interface was the design and layout, while the input modalities for operating various functions like navigation, media, comfort, and assistance functions remained unchanged. In light of the entry restrictions imposed by the government of the United States (US) in response to the COVID-19 pandemic, we were unable to include the US in study 2.

Table 1. Overview of the multi-study research approach.

Label	Market	Year
Study 1	Germany	2018
	China	2019
	US	2019
Study 2	Germany	2020
	China	2020
Study 3	Germany	2021
	China	2022
	US	2022

In each study, participants from each market interacted with the same vehicle. As benchmark studies aim at comparing products to those of competitors, we included three other reference vehicles in each market that were considered main competitors in the explicit market. Different vehicles were used as references in the different markets, as each market produces its own domestic brands. Currently, Chinese domestic brands, for

example, are not available in the US and are just at the doorstep of the European consumer market. Therefore, a strict separation between market and domestic brands within the scope of a real-world driving study seems inappropriate. Nevertheless, within the scope of the present work, we were not interested in comparing multiple vehicles but rather in comparing different user groups.

2.1. Participants

In study 1, $n = 102$ participants from Germany, China, and the US participated in the study. Study 2 included $n = 73$ participants from Germany and China. Study 3 included $n = 126$ participants from Germany, China, and the US. National subsamples were compared along several demographic variables (Table 2). The age difference was intentional since participants were recruited explicitly with respect to certain age criteria, e.g., the comparatively young average age of a Chinese new car buyer [24]. All participants were members of the public who were recruited through local field agencies and compensated for their participation. They did not receive any prior training. All participants were customers of premium vehicles whose current vehicle was not older than three years and not newer than six months. Thus, they had experience with novel digital user interfaces in general. In order to differentiate between premium-level and non-premium-level vehicles, we used brand reputation as a classification criterion. Furthermore, premium vehicles can be distinguished from non-premium vehicles by their higher price point [25].

Table 2. Demographics for study 1, study 2, and study 3.

Study	Market	n	Sex		Age	
			Female	Male	Mean (M)	Standard Deviation (SD)
1	Germany	30	3	27	54.0	11.2
	China	36	8	28	35.5	7.1
	US	36	12	24	39.5	8.3
	Total	102	23	89		
2	Germany	36	10	26	41.5	11.8
	China	37	9	28	35.2	7.9
	Total	73	19	54		
3	Germany	37	7	30	40.0	13.0
	China	39	9	30	35.0	7.2
	US	50	15	35	50.7	10.6
	Total	126	31	95		

2.2. Human–Machine Interface

The vehicles in the present work have typical solutions for modern interfaces. Thus, the insights generalize to a variety of modern vehicle interfaces. In each vehicle, interactions were available via tangible interfaces, i.e., remote controllers, hard keys, and steering wheel control switches, and via graphical user interfaces, i.e., an instrument cluster and a central information touchscreen. Figure 1 schematically illustrates the interior layout and highlights the relevant input modalities that will be analyzed in this work, including a touch display and remote-control element.

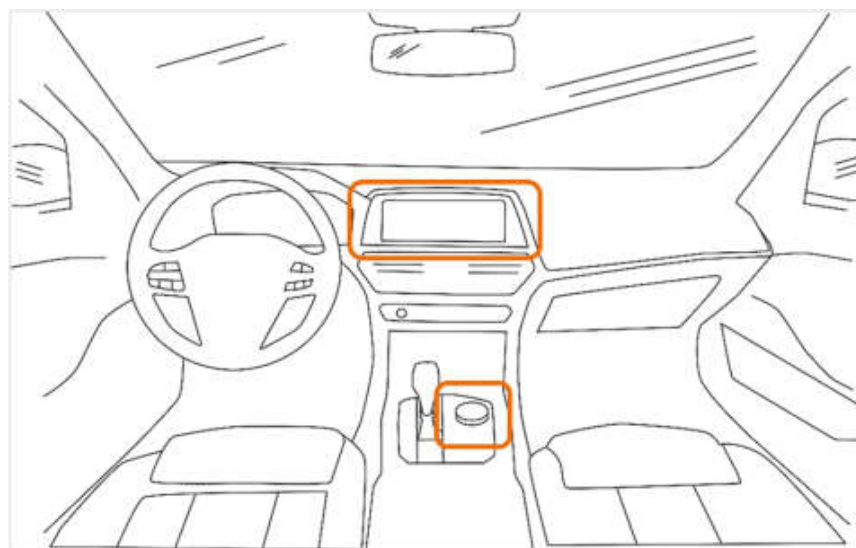


Figure 1. Layout of a modern HMI used in the studies including steering wheel control switches, hard keys, displays, and remote-control elements. Interaction modalities are highlighted in orange.

2.3. Material

2.3.1. Measurement of Satisfaction

Satisfaction is one of the three components of usability evaluation [10] and is considered an important evaluation criterion for HMIs [26]. Each participant was asked to rate the usability of the system using the system usability scale (SUS), which is considered the most widely used self-report measure of usability [27]. The SUS is a technology-agnostic tool and can therefore be used to assess the usability of a wide range of products, such as phones, IVISs [28], and automated driving [29]. The questionnaire consists of ten items, each to be answered on a five-point Likert scale from “strongly agree” to “strongly disagree”. Numerous studies have confirmed its excellent psychometric properties [30,31]. Moreover, prior research has shown the psychometric reliability and validity of the SUS in specific languages. Both the German and the Chinese translations seem to be capable of measuring the subjective usability of an infotainment system [32].

2.3.2. Measurement of Hedonic Qualities

The user experience questionnaire (UEQ) [33] was added to assess positive emotions and attitudes toward the human–machine interaction. The questionnaire covers usability aspects (efficiency, perspicuity, dependability) and UX aspects (originality, stimulation, novelty). The UEQ contains six scales with 26 items in total. The items have the form of a semantic differential, i.e., each item anchors with bipolar adjectives at each end, such as attractive versus unattractive. Prior research indicated acceptable to good reliability for all six subscales [34].

2.3.3. Measurement of Overall Evaluation

A widely used metric to rate the likelihood one would recommend a product, system, or service is the net promoter score (NPS). It is an aggregate-level measure that is derived from scores on a single survey item [35]. We included the NPS to provide a comprehensive global assessment, which takes into account multiple factors contributing to the user’s experience.

2.3.4. Measurement of Interaction Performance

There are two types of data (i.e., self-report versus behavior); therefore, we included both types in the studies. Besides reaction time and error rates as measures of speed and accuracy, observations by the experimenter can be a helpful tool to assess interaction

performance [36]. Experimenter ratings serve as one solution to the difficulty and extensive effort of collecting behavioral data [37,38]. The experimenter rated participants’ behavior on a five-point rating scale for user interaction success (Table 3) right after each UC during the experiment [39].

Table 3. Experimenter rating with label and description.

Category	Value	Description
No problem	1	<ul style="list-style-type: none"> No problem
Hesitation	2	<ul style="list-style-type: none"> Independent solution without errors But: hesitation, very conscious operating, and full concentration
Minor errors	3	<ul style="list-style-type: none"> Independent solution without or with minor errors, which were corrected confidently But: longer pauses for reflection Evaluation of potential operating steps
Massive errors	4	<ul style="list-style-type: none"> One or multiple errors Clearly impaired operation flow Excessive correction of errors No help from experimenter necessary
Help of experimenter	5	<ul style="list-style-type: none"> Multiple errors Massive errors require to restart task Help of experimenter necessary

Prior to the experimental procedure, each category had been adjusted specifically to each UC so that the generic categories were linked to specific descriptions of behavior for each UC. During the pilot tests, the experimenters discussed reasons for potential deviations in the ratings. The training aimed to support consistent coding during the study procedure.

Considering the number of UCs (see Table 4) and available modalities, analyses of the observational data were limited to two main modalities: the touchscreen and remote-control elements. Three UCs were selected from the main categories of navigation, media, and communication (e.g., UC 1 for navigation, UC 6 for media, UC 8 for communication). In their entirety, these three UCs provide a comprehensive overview of the interaction logic and design of a vehicle’s infotainment system.

Table 4. Overview of use cases for studies 1, 2, and 3.

Use Case Number	Task	Mode
1	Start navigation	P
2	Cancel navigation	P
3	View call list	P
4	Change volume/mute	D
5	Restaurant list	D
6	Skip radio station	D
7	Adjust temperature	D
8	Call contact	D
9	Play song	P
10 *	Send voice message	P

Note: all use cases were performed in each available modality; * only performed in study 3. The use cases of interest that have been evaluated in this work are presented in bold.

2.4. Study Design and Procedure

To compare the self-report and experimenter ratings between the user groups within each study, we conducted analyses with the market as a between-subjects factor. For the experimenter ratings, we further included the UCs (navigation-UC, media-UC, communication-UC) and modality (touch, remote) as within-subject factors.

The experiment began with participants signing a consent form and completing demographic questionnaires. The questionnaires used for the three cultural user groups were presented in the respective languages of the countries. After introducing the vehicle to the participants, they were asked to rate their first impression of the displays and operating modalities. Afterwards, the experimenter explained the tasks that should be performed. In each market, a corresponding native-speaking experimenter conducted the study. The studies followed the same UC set, which consisted of nine UCs (Table 4) in studies 1 and 2 and ten UCs in study 3. The experimenter sat in the passenger seat throughout the whole study and read the tasks out loud. The UCs were carried out in the given order: UCs 1, 2, 3, 9, and 10 were conducted in a parked situation, whereas the other UCs were conducted while driving. Each UC was performed with every available modality, but it was the participants' choice of which modality to use first. If participants felt unsafe conducting a UC using the touchscreen or the remote-control element while driving, they skipped the respective modality due to reasons of safety. The experimenter completed experimenter ratings after each UC during the experiment. After conducting a UC in each modality, the subsequent UC was carried out. The tasks were always started from the home screen of the IVIS. Participants completed the use cases while driving on a pre-defined route with a speed limit of 30 km/h. Driving safety had the highest priority; therefore, drivers were to refuse to complete UCs if they seemed too distracting. After conducting the UC set, the overall rating was executed in the vehicle by completing the SUS and the UEQ. Also, an in-vehicle evaluation of overall satisfaction and the voice assistant's visual, emotional, and functional value was collected. A final interview, including capturing the intention to recommend the system by the NPS, took place outside the vehicle. This resulted in a total length of about 60 min.

2.5. Statistical Procedure

Data from the questionnaires were analyzed using IBM SPSS Statistics Version 26, including, where appropriate [40], *t*-tests, univariate analysis of variance (ANOVA), univariate analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA), and multivariate analysis of covariance (MANCOVA). For ANCOVA and MANCOVA, we will report the adjusted means, which are calculated by taking into account the values of the covariates. When conducting multiple statistical tests, the risk of a type I error increases. Therefore, we employed appropriate correction techniques as a post hoc procedure to correct the family-wise error rate following the ANOVA [40,41]. For the analyses of the overall evaluation using the NPSs, the assumptions were violated. Thus, non-parametric tests, e.g., the Mann-Whitney U test and Kruskal-Wallis test, were calculated. As mentioned in Section 2.1, the age differences between the samples were intentional. However, age could still have a confounding impact on the effects. Therefore, the first step prior to the main analyses was to determine the correlation between the respective dependent variable (DV) and age. Where we found significant correlations, we chose a more complex approach and included age as a covariate in the analyses. Table 5 outlines pertinent constructs with their corresponding scales and subscales, the studies in which they have been included, and the analyses employed.

Table 5. Overview of relevant constructs, their respective data types, methods, and analyses.

Data Type	Construct	Method	Subscales	Source	Studies	Analyses Applied
Self-report measures	Satisfaction	SUS	Attractiveness	Brooke [27]	1, 2, 3	<i>t</i> -test, ANOVA
	Pragmatic and hedonic qualities	UEQ	Perspiciuity Efficiency Dependability Stimulation Novelty	Laugwitz et al. [33]	2, 3	MANCOVA
	Overall evaluation	NPS		Reichheld [42]	1, 2, 3	Mann–Whitney U test, Kruskal–Wallis test
Observational measures	Interaction performance	Experimenter rating	UC (Navi, Media, Communication) Modality (Touch, Remote)	Naujoks et al. [36]	1, 2, 3	Mixed within-between ANCOVA

3. Results

3.1. Results: Study 1

3.1.1. SUS

There was no significant correlation between age and the DV, $r = 0.08$, $p = 0.441$; therefore, there was no indication of a confounding influence of age, and we did not include age as a covariate. A univariate ANOVA was conducted to assess the effect of the between-subjects factor market (Germany, China, US) on usability ratings. The Levene test indicated equal variances ($p = 0.890$). The ANOVA showed no significant effect of the market on the SUS scores, $F(2,99) = 1.07$, $p = 0.346$, $\eta_p^2 = 0.02$, indicating that participants' satisfaction ratings did not differ significantly between the markets (Table 6).

Table 6. Descriptive statistics (i.e., *M* and *SD*) for the SUS scores in study 1.

Market	<i>M</i>	<i>SD</i>
Germany	72.08	14.26
China	73.89	19.50
US	78.89	21.22

3.1.2. NPS

There was no significant correlation between age and the DV, $r = 0.08$, $p = 0.489$. The NPS was only conducted in China and the US; therefore, only these two markets are included in the analysis. Since the assumptions of normal distribution and equality of variances were violated, a non-parametric approach was chosen. A Mann–Whitney U test revealed that the Chinese and American participants differed significantly in their intention to recommend the system ($z = -3.28$, $p = 0.001$, $r = 0.39$). Chinese participants ($Mdn = 8.00$) were less likely to recommend the system than American participants ($Mdn = 9.50$).

3.1.3. Experimenter Ratings

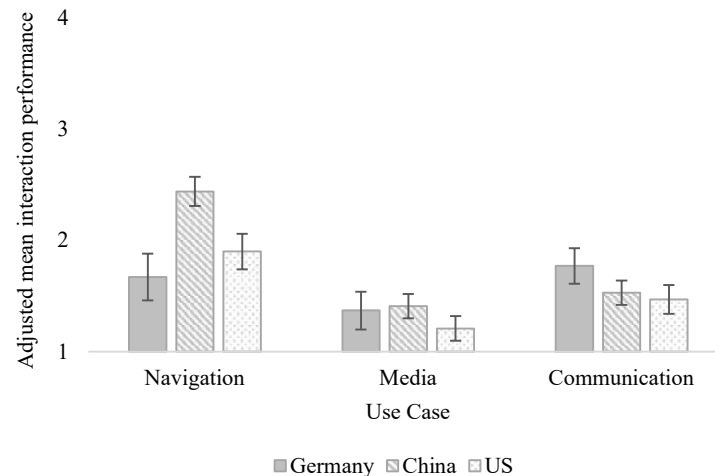
Since we found significant correlations between age and the DV (highest $r = 0.44$, $p < 0.001$), we calculated a mixed between-within ANCOVA with *market* (Germany, China, US) as the between-subjects factor and the *UC* (navigation-UC, media-UC, communication-UC) and *modality* (touch, remote control) as the within-subject factors. We further included *age* as a covariate. The assumptions of sphericity were not violated (all $\chi^2(2) < 1.85$, $p > 0.380$). The results of the $2 \times 3 \times 3$ ANCOVA can be found in Table 7.

Table 7. Inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the mixed between-within ANCOVA for experimenter ratings in study 1.

Effect	$df1$	$df2$	F	p	η_p^2
Market	2	94	2.45	0.092	0.50
Age	1	94	12.30	<0.001	0.12
UC	2	188	1.27	0.283	0.01
Modality	1	94	0.92	0.340	0.01
UC \times Market	4	188	2.53	0.042	0.05
UC \times Age	2	188	0.71	0.495	0.007
Modality \times Market	2	94	2.07	0.132	0.04
Modality \times Age	1	94	0.02	0.904	0.00
UC \times Modality	2	188	1.42	0.245	0.02
UC \times Modality \times Market	4	188	1.88	0.115	0.04
UC \times Modality \times Age	2	188	0.17	0.841	0.002

Note: Significant effects are in bold.

Interaction UC and Market. Pairwise comparisons using Bonferroni correction revealed a significant difference between German and Chinese participants for the navigation UC, $t(2) = 2.47$, $p = 0.044$, $d = 0.67$, indicating that the performance was significantly worse in the Chinese ($M = 2.44$, $SE = 0.17$) compared to the German sample ($M = 1.70$, $SE = 0.21$). According to Cohen (1988) [43], this is a medium-sized effect. While German participants showed the most difficulties initiating the phone call, Chinese and American participants had the most problems starting the navigation. To examine the interaction in more detail, Figure 2 presents the data in a graphical format. The examination revealed a disordinal interaction, meaning that the differences between the two markets can only be interpreted in combination with the levels of the UC factor.

**Figure 2.** Adjusted mean interaction performance scores for the three UCs split by market. Note: higher values indicate worse performance, i.e., having more trouble completing the tasks. The error bars represent the standard errors (SEs).

3.2. Results: Study 2

3.2.1. SUS

First, we checked for correlations between age and the DV. There was no significant correlation: $r = 0.21$, $p = 0.078$. We calculated a t -test for independent samples to examine if German and Chinese participants differ in satisfaction ratings. Since the Levene test was significant ($F(1,71) = 13.98$, $p < 0.001$), we could not assume equality of variance. Results of the t -test revealed a significant difference between the two markets: $t(71) = 4.08$, $p < 0.001$. German participants rated the interface significantly higher ($M = 78.68$, $SD = 11.32$) than Chinese participants ($M = 61.96$, $SD = 21.91$).

3.2.2. UEQ

A MANCOVA was calculated with *market* as the between-subjects factor. The six subscales of the UEQ, representing the construct of UX, formed the DVs. *Age* was included as a covariate because prior analyses revealed significant correlations between age and UEQ subscale *perspicuity* ($r = 0.24, p = 0.044$), *efficiency* ($r = 0.23, p = 0.048$), and *stimulation* ($r = 0.23, p = 0.046$). The MANCOVA revealed a significant effect of *market* on the UEQ subscales: $F(6,65) = 5.48, p < 0.001, \eta_p^2 = 0.34$, Wilk’s $\Lambda = 0.66$. There was no significant effect of age: $F(6,65) = 0.48, p = 0.820, \eta_p^2 = 0.04$, Wilk’s $\Lambda = 0.96$. Subsequent ANCOVAs showed significant differences between the German and Chinese samples in each subscale except for the subscale *novelty* (Table 8).

Table 8. Descriptive statistics (i.e., adjusted *M*, *SE*) for the UEQ subscales by market for study 2, and inferential statistics (i.e., *df1*, *df2*, *F*-, *p*-, and η_p^2 -values) for the univariate ANOVAs.

UEQ Scales	Market	Adjusted <i>M</i>	<i>SE</i>	<i>df1</i>	<i>df2</i>	<i>F</i>	<i>p</i>	η_p^2
Attractiveness	Germany	2.00	0.19	1	70	15.67	<0.001	0.18
	China	0.94	0.18					
Perspicuity	Germany	1.69	0.18	1	70	12.34	<0.001	0.15
	China	0.76	0.18					
Efficiency	Germany	1.62	0.19	1	70	6.68	0.012	0.09
	China	0.90	0.19					
Dependability	Germany	1.97	0.18	1	70	15.05	<0.001	0.18
	China	0.97	0.18					
Stimulation	Germany	1.67	0.19	1	70	17.68	<0.001	0.20
	China	0.54	0.18					
Novelty	Germany	0.81	0.23	1	70	2.52	0.117	0.04
	China	0.30	0.22					

Note: Significant effects are in bold.

According to Laugwitz and colleagues [33], certain subscales of the UEQ can be grouped into pragmatic qualities, i.e., *perspicuity*, *efficiency*, and *dependability*, and hedonic qualities, i.e., *stimulation* and *originality*. While pragmatic qualities describe task-related quality aspects, hedonic qualities represent non-task-related quality aspects. Furthermore, there are items that directly measure the perceived attractiveness. Figure 3 displays the distribution of pragmatic and hedonic quality aspects and perceived attractiveness in the German and Chinese samples.

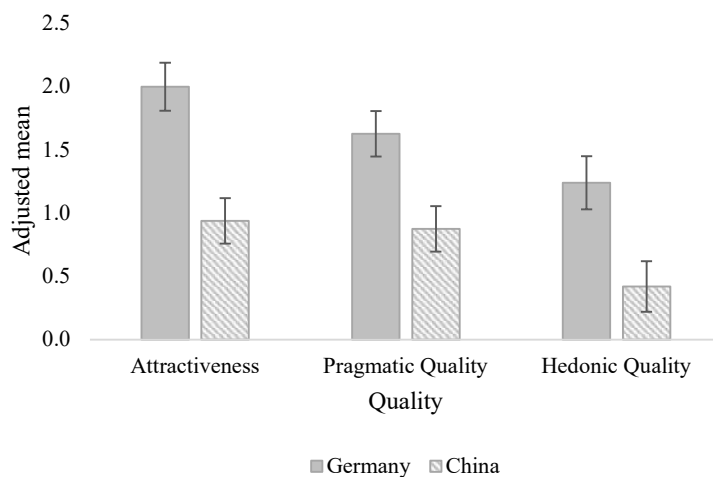


Figure 3. Distribution of pragmatic and hedonic qualities of participants from the German and Chinese markets for study 2. Note: the error bars represent the standard errors (*SEs*).

3.2.3. NPS

There was no significant correlation between age and the DV: $r = 0.16$, $p = 0.177$. Since the assumptions of normal distribution and equality of variances were violated, a non-parametric approach was chosen. A Mann–Whitney U test revealed that the German and Chinese participants differed significantly in their intention to recommend the system to others ($z = -5.22$, $p < 0.001$, $r = 0.61$). Chinese participants ($Mdn = 6.00$) were less likely to recommend the system than German participants ($Mdn = 8.50$).

3.2.4. Experimenter Ratings

A mixed between-within ANCOVA was calculated with *market* as a between-subjects factor and *UC* and *modality* as within-subject factors. We further included *age* as a covariate because we found a significant correlation between age and performance in communication UC using remote control ($r = 0.28$, $p = 0.017$). Therefore, the reported effects are adjusted for the influence of age. The assumption of sphericity was violated for the interaction of the factors UC and modality: $\chi^2(2) = 17.61$, $p < 0.001$; therefore, the Greenhouse–Geisser corrected values are reported for the interaction. The results of the $2 \times 2 \times 3$ ANCOVA can be found in Table 9.

Table 9. Inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the mixed between-within ANOVA for experimenter ratings in study 2.

Effect	$df1$	$df2$	F	p	η_p^2
Market	1	67	12.79	<0.001	0.16
Age	1	67	2.30	0.134	0.03
UC	2	134	4.60	0.012	0.06
Modality	1	67	1.47	0.230	0.02
UC \times Market	2	134	9.87	<0.001	0.13
UC \times Age	2	134	2.62	0.077	0.04
Modality \times Market	1	67	6.33	0.014	0.09
Modality \times Age	1	67	0.01	0.916	<0.01
UC \times Modality	1.62	108.57	2.47	0.100	0.04
UC \times Modality \times Market	2	134	4.58	0.012	0.06
UC \times Modality \times Age	2	134	0.64	0.529	0.01

Note: Significant effects are in bold.

Market. Experimenter ratings were significantly higher in the Chinese sample ($M = 1.80$, $SE = 0.08$) than in the German sample ($M = 1.40$, $SE = 0.08$), indicating that performance was worse for Chinese participants.

Use Case. Pairwise comparisons using Bonferroni–Holm correction showed significant differences between the navigation UC and the media UC ($t(298) = 7.58$, $p < 0.001$, $d = 0.89$) and between the navigation UC and the communication UC ($t(299) = 6.75$, $p < 0.001$, $d = 0.79$). Experimenter ratings were significantly higher for the navigation UC ($M = 2.18$, $SE = 0.10$) than for the media UC ($M = 1.26$, $SE = 0.07$) and communication UC ($M = 1.36$, $SE = 0.08$). This result indicates that performance was the worst in the navigation UC across the two countries.

Interaction: UC and Market. Results revealed a significant interaction between the UC and market, which indicates that the effect of the UC is not equal for the two samples. Across both markets, experimenter ratings were highest for the navigation UC. However, in the Chinese sample, the difference in performance between the navigation UC and the media and communication UC was greater than the difference in the German sample. Detailed descriptive data on experimenter ratings and a figure displaying the interaction are provided in Appendix A (Table A1, Figure A1). Both main effects can be interpreted globally, as it was an ordinal interaction.

Interaction: Modality and Market. Results showed a significant ordinal interaction between modality and market. Across all three UCs, participants' performance was worse when using remote control to operate the system than it was when using touch. However,

we could not observe a main effect of modality since there was no significant difference between the two markets in the touch modality: $t(433) = 1.27, p = 0.200$. Only when using remote control to operate the system did the German and Chinese participants differ significantly in their performance: $t(433) = 3.61, p < 0.001$. Detailed descriptive data and a figure are provided in Appendix A (Table A1, Figure A1).

Interaction: UC, Modality, and Market. Inferential analyses revealed a significant three-way interaction. To further examine the interaction, we conducted pairwise comparisons using Bonferroni correction. Figure 4 displays the ratings split by modality to promote a better understanding of the higher-order interaction. The findings indicate that the $UC \times market$ interaction differs for one level of the factor *modality* to another level of the factor: Chinese participants had more difficulty starting the navigation compared to German participants, but solely when using remote control ($t(71) = 4.42, p < 0.001$) and not using touch control ($t(71) = 1.55, p = 0.130$). There were no significant differences between the German and Chinese participants for the other two UCs, i.e., media and communication, neither when using touch ($t(71) = 0.31, p = 0.733$) nor when using remote control ($t(71) = 0.30, p = 0.760$). For that reason, there was no main effect of *modality*. The modality exclusively impacted performance depending on the UC, resulting in a three-way interaction. Further, the $modality \times market$ interaction differed for the levels of factor *UC*: there were significant differences between the German and Chinese samples for operating the system using remote control compared to touch, but only for the navigation UC.

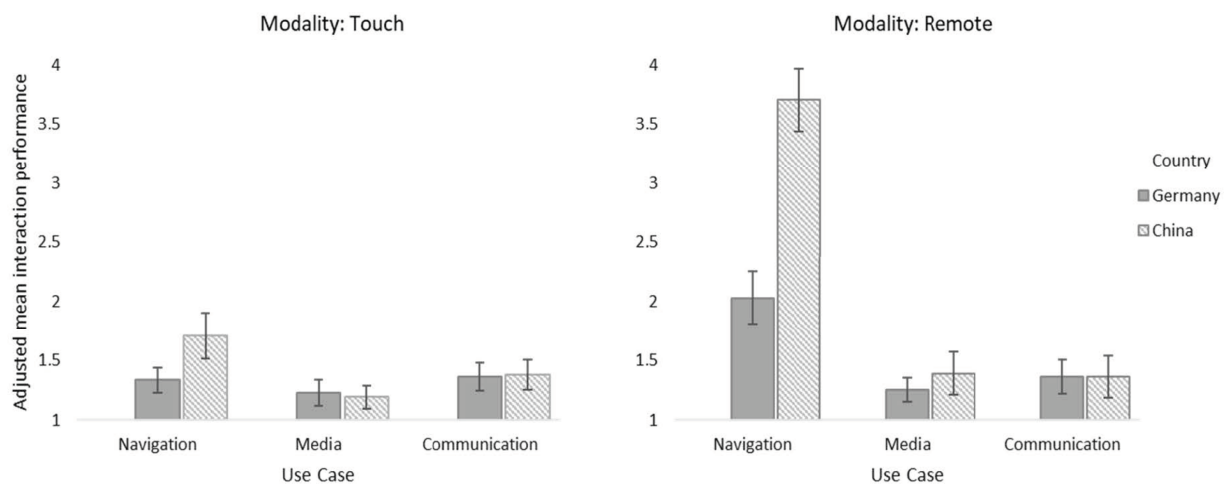


Figure 4. Adjusted mean interaction performance scores of German and Chinese participants for the three UCs split by modality in study 2. Note: higher values indicate worse performance, i.e., having more trouble completing the tasks. The error bars represent the standard errors (SEs).

3.3. Results: Study 3

3.3.1. SUS

There was no significant correlation between age and the DV: $r = 0.04, p = 0.663$. A univariate ANOVA was conducted to assess the effect of the between-subjects factor *market* (Germany, China, US) on usability ratings. The Levene test indicated unequal variances ($p = 0.005$). The ANOVA revealed a significant effect of the market on the SUS scores: $F(2,123) = 4.25, p = 0.016, \eta_p^2 = 0.07$. Post hoc t -tests using Bonferroni–Holm correction showed a significant difference between the German and Chinese participants ($t(62.44) = 2.97, p = 0.004, d = 0.65$), with participants from the German sample giving significantly higher scores than those from the Chinese sample. According to Cohen (1988) [43], this is a medium-sized effect. After correction, no additional statistically significant effects were observed. Table 10 shows the descriptive data.

Table 10. Descriptive statistics (i.e., *M* and *SD*) for the SUS in study 3.

Market	<i>M</i>	<i>SD</i>
Germany	79.73	12.77
China	67.76	21.48
US	71.65	18.94

3.3.2. UEQ

A MANCOVA was calculated with *market* as the between-subjects factor. The six subscales of the UEQ, representing the construct of UX, formed the DVs. *Age* was included as a covariate because prior analyses revealed significant correlations between age and the UEQ subscales *attractiveness* ($r = 0.22, p = 0.015$), *efficiency* ($r = 0.22, p = 0.015$), *dependability* ($r = 0.22, p = 0.015$), *stimulation* ($r = 0.26, p = 0.002$), and *novelty* ($r = 0.28, p = 0.002$). One participant was excluded because they gave the same answer for each item, resulting in a total of $N = 125$ completed questionnaires. There was no significant effect of *age* ($F(6,111) = 1.67, p = 0.135, \eta_p^2 = 0.08, \text{Wilk's } \Lambda = 0.92$), and no significant effect of *market* on the UEQ subscales ($F(12,222) = 1.72, p = 0.064, \eta_p^2 = 0.09, \text{Wilk's } \Lambda = 0.84$). The descriptive data, including adjusted means and *SE*, can be found in Table 11.

Table 11. Descriptive data (i.e., adjusted *M* and *SE*) for the six UEQ subscales grouped by market.

UEQ Subscale	Market	Adjusted <i>M</i>	<i>SE</i>
Attractiveness	Germany	2.02	0.16
	China	1.34	0.17
	US	1.74	0.17
Perspicuity	Germany	1.71	0.18
	China	1.34	0.19
	US	1.09	0.19
Efficiency	Germany	1.70	0.18
	China	1.22	0.18
	US	1.45	0.18
Dependability	Germany	1.82	0.17
	China	1.31	0.17
	US	1.40	0.17
Stimulation	Germany	1.67	0.17
	China	1.20	0.17
	US	1.56	0.17
Novelty	Germany	1.52	0.18
	China	1.04	0.18
	US	1.19	0.18

Figure 5 displays the distribution of pragmatic and hedonic quality aspects and perceived attractiveness in the German, Chinese, and American samples.

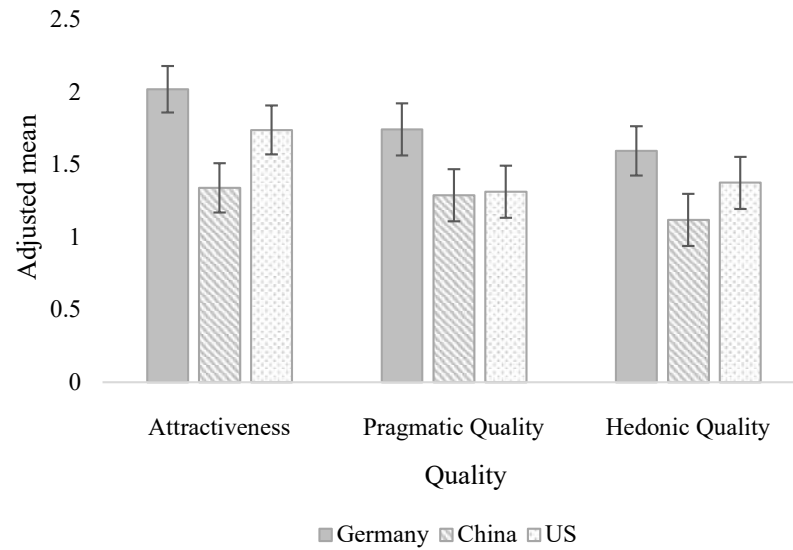


Figure 5. Distribution of pragmatic and hedonic qualities of participants from Germany, China, and the US for study 3. Note: the error bars represent the standard errors (SEs).

3.3.3. NPS

There was no significant correlation between age and the DV: $r = 0.09, p = 0.318$. Both the assumption of normal distribution and equality of variances were violated; hence, a non-parametric procedure was chosen. A Kruskal–Wallis test revealed a significant difference in the intention to recommend the system between the three markets ($H(2) = 12.51, p = 0.002$). A post hoc test, the Dunn–Bonferroni test, showed that the probability of Chinese participants ($Mdn = 6.00$) recommending the system was significantly lower compared to German participants ($Mdn = 8.00, z = 2.87, p = 0.012, d = 0.66$) and American participants ($Mdn = 8.00, z = -3.26, p = 0.003, d = 0.96$).

3.3.4. Experimenter Ratings

Touch. To compare experimenter ratings within the touch modality, a mixed between-within ANCOVA was calculated with *market* (Germany, China, US) as a between-subjects factor and *UC* as a within-subject factor. We included *age* as a covariate because prior analyses revealed significant correlations between age and the DV for the navigation UC ($r = 0.33, p < 0.001$) and the communication UC ($r = 0.35, p < 0.001$). The assumption of sphericity was violated for UC ($\chi^2(2) = 27.65, p < 0.001$); therefore, Greenhouse–Geisser corrected values are reported for this effect. The results of the 3×3 ANCOVA are displayed in Table 12.

Table 12. Inferential statistics (i.e., $df1, df2, F, p$ -, and η_p^2 -values) for the mixed between-within ANCOVA for experimenter ratings in study 3 (touch modality).

Effect	<i>df1</i>	<i>df2</i>	<i>F</i>	<i>p</i>	η_p^2
Market	2	115	11.10	<0.001	0.16
Age	1	115	22.47	<0.001	0.16
UC	1.65	189.25	5.48	0.005	0.05
UC \times Market	4	230	7.08	<0.001	0.11
UC \times Age	2	230	10.76	<0.001	0.09

Note: Significant effects are in bold.

Market. Pairwise comparisons using Bonferroni correction revealed a significant difference in interaction performance between German and Chinese participants ($t(73) = 3.9, p < 0.001, d = 0.90$) and between German and American participants ($t(79) = 3.7, p < 0.001, d = 0.83$). The findings indicate that German participants had more difficulty operating

the system using touch across all three UCs ($M = 1.60$, $SE = 0.07$) compared to Chinese ($M = 1.21$, $SE = 0.07$) and American participants ($M = 1.23$, $SE = 0.07$).

Use Case. As pairwise comparisons using Bonferroni correction showed, there was a significant difference in performance between the navigation UC and media UC ($t(236) = 5.44$, $p < 0.001$, $d = 0.71$) and the navigation UC and communication UC ($t(236) = 3.18$, $p = 0.006$, $d = 0.41$). The participants had significantly higher scores in the navigation UC ($M = 1.62$, $SE = 0.09$) than in the communication ($M = 1.14$, $SE = 0.04$) and media UCs ($M = 1.27$, $SE = 0.06$), indicating the worst performance was in the navigation UC.

Interaction: UC \times Market. While German ($M = 2.27$, $SE = 0.16$) and Chinese participants ($M = 1.40$, $SE = 0.16$) struggled the most with starting the navigation, participants from the US had the most trouble with calling a contact ($M = 1.36$, $SE = 0.11$). A further examination of this interaction revealed a disordinal interaction, indicating that neither of these main effects can be interpreted globally. Differences between the three countries can only be interpreted in combination with the three UCs and vice versa. Detailed descriptive data are provided in Appendix B (Table A2, Figure A2).

Remote control. Because there were no experimenter ratings for the remote-control modality for the German sample in study 3, only the ratings from the Chinese and American samples were compared in the analysis. We found significant correlations between age and the DV (all $r > 0.22$, all $p < 0.042$) and included *age* as a covariate. The assumption of sphericity for the 2×3 ANCOVA with *market* as a between-subjects factor, *UC* as a within-subject factor, and *age* as a covariate was violated for UC: $\chi^2(2) = 46.01$, $p < 0.001$. Therefore, the Greenhouse–Geisser corrected values are reported. The ANCOVA revealed that only the covariate had a significant effect on interaction performance: $F(1,79) = 5.82$, $p = 0.018$, $\eta_p^2 = 0.07$. The remaining inferential statistics can be found in Appendix C (Table A3).

3.4. Overview of Satisfaction and Interaction Performance for German and Chinese Users across the Three Studies

In the following section, we will visualize the data from the three studies. No statistical analysis or inclusion of temporal factors (study 1, 2, 3) is conducted, as this is not a true between-design and different participants were involved in each year, with a different interface tested in study 3. The visualization serves the purpose of comparing self-report and observational measures. We illustrate satisfaction ratings and interaction performance over the course of the three studies for Germany and China (Figures 6 and 7). The figures do not include data from the US sample since study 2 was conducted in Germany and China only.

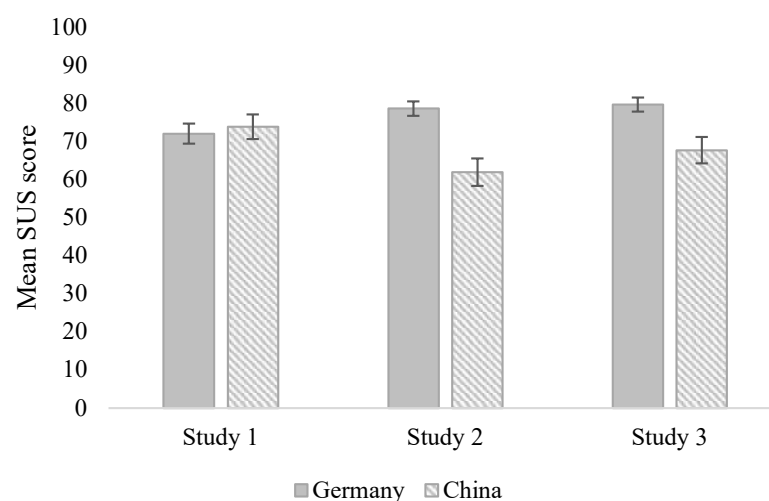


Figure 6. Satisfaction ratings over the course of time for German and Chinese users. Note: higher ratings indicate higher satisfaction ratings. The error bars represent the standard errors (SEs).

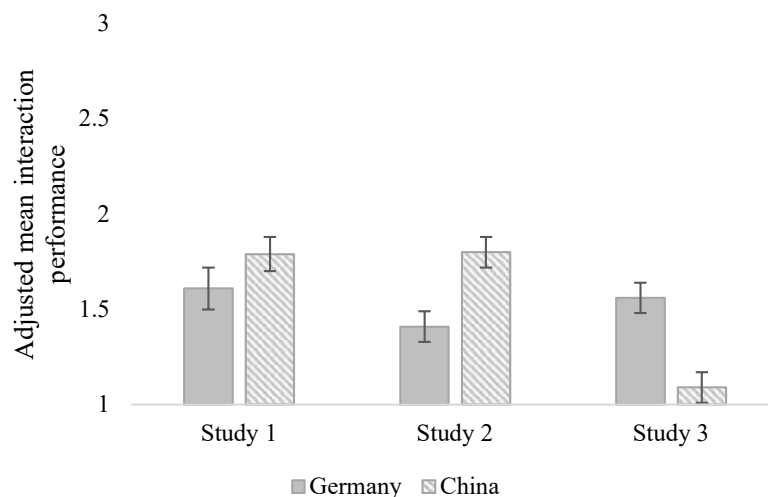


Figure 7. Interaction performance over the course of time for German and Chinese users. Note: higher values indicate worse performance, i.e., having more trouble completing the tasks. The error bars represent the standard errors (SEs).

4. General Discussion

This work outlined the findings of a series of international studies that investigated how users from different markets interact with the same in-vehicle infotainment system. In real driving scenarios, $N = 301$ participants from Germany, China, and the US completed several tasks using different interaction modalities and rated the system after the interaction. We identified differences in interaction performance, satisfaction, hedonic qualities, and overall ratings between the user groups. The following paragraphs discuss the present findings, considering study limitations, and point towards future research opportunities for benchmark studies in the automotive context.

4.1. Differences in Satisfaction

The results of the work at hand suggest the existence of differences between different user groups in satisfaction ratings. German participants reported the highest satisfaction with the interface, while the ratings were lowest in China (Figure 6). German users favored the overall usability and satisfaction of the interface compared to Chinese users. In terms of descriptive categorization, the scale provides a framework for identifying areas in need of action. The scores obtained from the questionnaires can be classified into different categories, ranging from the “worst imaginable” to the “best imaginable”. In our study, the scores ranged from “okay” to “good”. It is noteworthy that only the ratings of Chinese participants fell into the category of “okay”, which, according to the scale, implies a need for action and improvement. A German OEM developed the vehicle that has been evaluated in the studies. As Lindgren and colleagues [16] pointed out, HMI design and development often focus primarily on the needs and preferences of drivers from the Western market. Needs and expectations, in turn, can impact usability evaluations [44]. Mehler and colleagues [45] designed two interfaces: one adapted to the preferences of a Chinese sample and the other one adapted to the preferences of a German sample. The authors found that Chinese participants preferred the Chinese HMI over the German HMI, and vice versa. The interface used in this study was rather adapted to the needs of the Western market since it was developed by a German OEM and therefore received higher satisfaction from participants living and driving in the Western world. In a study by Khan and Williams [19], the authors compared two user groups from India and the United Kingdom and found some discrimination between the groups regarding usability factors, including satisfaction. They found that each group had a preferred approach to using information for problem-solving, which was determined by their respective cultural backgrounds. This determination subsequently impacts the expected visual presentation of information in a

vehicle HMI system. When there is a mismatch between the user's preferred approach and the visual information display, users tend to express dissatisfaction. Assuming there was a mismatch between the expectations of Chinese users and the displayed information in the studies, this mismatch could account for the lower satisfaction ratings. Through this, the necessity for an adapted market-specific HMI solution for the Eastern market becomes striking.

The findings of our multi-study approach revealed differences in satisfaction between German and Chinese users. However, when comparing with US users, the observed differences remained primarily descriptive in nature. A potential explanation could be found in models that address culture and cultural differences. Cultural frameworks aim to describe and define cross-cultural differences. Hofstede, for example, defines culture as the "collective programming of the mind that distinguishes the members of one group or category of people from others" [46] (p. 3). He describes culture along six dimensions and identifies differences between Germany, China, and the US [47]. Another cultural model by Trompenaars and Hampden-Turner [48] classifies culture into seven dimensions. The authors assume that differences between Germany, China, and the US occur in several dimensions, such as the perception of time. Nonetheless, our data did not reveal significant results for comparing satisfaction ratings with US users. This might be due to the methodology of this study. It is possible that the effects remained of a descriptive nature due to the given sample size. The effects are small in terms of the conventions introduced by Cohen [43]. To detect those small effects, a larger sample size would have been necessary. Conversely, this also underscores the practical significance of the effects that reached statistical significance in the between-subjects design with approximately 36 observations per cell (see Table 2). While the effect sizes were small, these effects emerged even with our relatively modest sample size. We selected medium-sized samples per market with the aim of detecting medium-sized effects. The fact that the apparently small effects were already significant given the present sample size can be considered meaningful.

There were no significant differences in satisfaction ratings in study 1 (Table 6). Each user group experienced a satisfactory interaction with the system. In the following study, one year later, differences occurred between German and Chinese users. Chang and colleagues [49] demonstrated that product preferences can be influenced by preconceived judgments due to the user's evaluative frame of reference. Those sequential effects might occur in the context of automotive benchmarking as well. Fundamental changes with switches to digital UI are developing rapidly. While the same UI was evaluated in study 1 and study 2, different reference vehicles were incorporated in these two studies. Overall, Chinese consumers showed that they appreciated digital innovations more than German consumers. They may have anticipated novel features, but the current UI failed to meet their expectations, resulting in diminished satisfaction ratings.

4.2. Differences in Hedonic Qualities

By means of a literature review, Souza and Bernardes [50] showed that cultural differences affect UX in several domains, such as product development, systems, or games. Our results of the UEQ may extend this insight to the automotive context: the user groups from Germany and China differed in their ratings of the UI's pragmatic and hedonic quality aspects in study 2 (Table 8). For each of the six UEQ subscales, the scores of the Chinese participants were below average, while the scores of the German participants were considered good or even excellent [51], with one exception: both user groups gave scores below average in the subscale *novelty*. This might indicate that both groups considered the interface to be rather conservative and conventional rather than innovative and inventive. In study 3 (Table 11), the differences were not significant. However, the descriptive data suggest a pattern of lower scores from Chinese users compared to the other two user groups. It is striking that, overall, the participants from China gave lower ratings than the participants from Germany and the US (Figures 3 and 5), which might lead to the assumption of different response tendencies in the Chinese sample. Chinese participants seemed to be

less euphoric when interacting with the system. One explanation might be again found in a cultural model: the sixth dimension, *indulgence vs. restraint*, of Hofstede's model [46] describes how members of a cultural group enjoy or suppress the gratification of natural human desires related to enjoying life and having fun. Sogemeier and colleagues [52] linked these cultural dimensions to usability criteria to gain a deeper comprehension of why the same products are rated differently in different markets. Considering cultural differences in this dimension to explain our results seems reasonable, as indulgence tends to prevail in Western Europe while restraint prevails in Asia. Less enthusiastic ratings in the Chinese sample might be rooted in more pessimistic attitudes that characterize restrained societies [46]. Indulgent cultures obtain more optimistic attitudes and express feelings of happiness and enjoyment more freely. This might account for the generally low ratings in the Chinese sample compared to the more positive ratings in the German and US samples.

Another dimension established by Hofstede [53], *uncertainty avoidance (UA)*, describes the extent to which members of a cultural group feel threatened by ambiguous situations. China has a low score on UA, while Germany rates this dimension highly. In weak UA cultures, people strive for novelty and are excited to try new things [54]. Chinese car manufacturers implement new features in their HMI, such as bigger screens, the integration of WeChat, or more exciting voice assistants. If new features are not being perceived as exciting or novel, hedonic qualities of the interaction might be missing.

4.3. Differences in Overall Ratings

The probability of recommending the vehicle to another person yielded significant differences between the user groups in each study. The results show that Chinese users were less likely to recommend the system to others than German and US users. These results corroborate the findings outlined above for UX measures in the sense that Chinese customers are somewhat more skeptical and less likely to be enthusiastic about vehicle technology.

4.4. Differences in Interaction Performance

Experimenter ratings were included as a measure of behavioral data to quantify interaction performance. The participants' age differences might have accounted for differences in interaction performance since age affects, among others, reaction time [55], computer-based task performance [56], and the usage of in-car information systems while driving [57]. Lerner and colleagues [58] further added that older drivers are less likely to engage in non-driving related tasks while driving compared to their younger or middle-aged counterparts. We therefore considered age's confounding influence on interaction performance in our analyses. The findings show that age should not be neglected as it can also explain observed differences to some extent (see Table 7, Table 12, and Table A3). Yet, these age effects may not be due to age per se, i.e., chronological age, but rather to age-related processes, such as a lack of experience with digital systems [59]. Hence, forthcoming research should inquire directly about the participants' familiarity and experience with digital systems. But age is not the only influencing factor. When eliminating the effect of age, participants from different countries still differed in their performance. Thus, the observed differences in performance go beyond age differences.

The results indicate that, across the three user groups, the navigation UC was the most difficult one to perform (in study 1, see Figure 2; in study 2, see Figure 4; in study 3, see Section 3.3.4). Experimenter ratings were highest for the navigation UC in each study, followed by the communication UC and the media UC. This finding supports considerations regarding the study design. The intent of the study was to implement a wide spectrum of UCs, with some being easier to perform than others. Although participants consistently experienced the greatest difficulty in the navigation UC, the interaction effects, as demonstrated in study 1 (refer to Table 7), study 2 (refer to Table 9), and study 3 (refer to Table 12), indicate that different user groups perceive distinct UCs to possess various

degrees of ease or difficulty. For instance, compared to German users, Chinese users perceived navigation as significantly more challenging than the other UCs (Figure 4).

With the results at hand, we cannot definitively conclude whether one operating method led to better or worse performance. In study 2, as illustrated in Figure 4, the interaction between market and modality showed that Chinese participants encountered more difficulty in completing tasks using the remote-control element compared to the touchscreen. The remaining results from study 1 and study 3 did not exhibit a distinct pattern.

The differences in experimenter ratings between the three user groups were ambiguous. As shown in Figures 6 and 7, satisfaction ratings assessed through the SUS did not consistently reflect measures of interaction performance. Only in study 2 did we observe a fit between self-report and observational measures, i.e., German participants showed better performance and rated the system higher. In study 3, however, even though Chinese participants were better at performing the UCs compared to German participants, they rated the interaction as significantly less satisfactory. Thus, the data at hand did not point towards a *reliable* predictive value of preference for performance. Roberts and colleagues [60] acknowledged a dissociation between preference and performance in a study that investigated the usability of metro maps: reaction times and error rates did not match self-reported satisfaction. Knapp [61] found that the performance of a Chinese user group was negatively affected if the system was based on a different user group's mental model, but even though the performance was worse in one experimental condition, participants did not perceive this condition as less attractive. Due to this gap, the author concludes that it is necessary to include a wide range of measures. In a literature review, Pettersson and colleagues [15] also concluded that the use of triangulation, i.e., including several methods in UX evaluation, is beneficial. Law and colleagues [62] stated that including quantitative measures while excluding qualitative measures and vice versa may lead to wrong implications. In our case, preference measures did not predict performance, and vice versa. Thus, collecting both self-report and observational data is essential to drawing a holistic picture of human-machine interactions. In addition, instead of relying on a single observational measure, it might be beneficial for researchers to include a broader array of methods, such as error rates, time on task, and reaction times. Likewise, Forster and colleagues [63] recommend a multi-method approach to provide superior insights into the quality of an HMI. Furthermore, performance indicators were aggregated over several UCs to generate a single measure for interaction performance. The data suggest that cutting down the number of variables for the sake of parsimony seems inappropriate due to an essential loss of information. An analysis of interaction on a more granular level, i.e., analyzing each UC individually, may provide more detailed information about the human-machine interaction to reveal further potential for the development of HMI. The results emphasize the need to include both self-reported preference measures and observational performance measures.

4.5. Limitations and Future Research

The present work comes with a few limitations. The studies were conducted over a period of approximately four years, and different interviewers guided the studies, which could have influenced objectivity. However, the studies followed a strict manual, and the interviewers received comprehensive training in advance to enable comparability. The questionnaires were given in the participants' respective languages, which may have impacted the results. Another limitation is that participants interacted with different reference vehicles in each study. The interaction with a previous product can foster sequence effects [49], which means that the evaluation of an object relates to what a participant has seen and experienced before. Those interaction effects between stimuli that are presented in successive order might occur in the context of automotive benchmarking, too. However, to avoid sequence effects in the present studies, the sequence of vehicles was permuted per participant [64]. Moreover, different reference vehicles in different markets were intended

for the sake of external validity. Today's domestic Chinese vehicle brands are not available in the US and are just at the doorstep of entering the European consumer market. We chose a real-world driving study and ensured an externally valid setting by respecting the fact that the availability of Eastern domestic brands in the Western market will still require time. In each market, we compared the newest HMI by a German OEM to the newest HMI by local competitors. This way, we provided a comparable framework by considering the technological progress and degree of novelty within each market. The majority of participants in our sample were male, which limits the generalizability of our findings. However, the participants were recruited with specific criteria regarding sex and age to represent a targeted customer group. Future research might include a more diverse sample to test the generalizability of the findings to a broader population.

The results of the work at hand highlight the potential of benchmark studies that enable the comparison of products such as IVISs in different international contexts. Future benchmark studies should also include ADSs to identify differences between user groups and to derive implications for future research and automotive HMI design.

In conclusion, from the results, a lot of valuable information can be summarized. There are differences between the three user groups, both in self-reporting and in behavioral measures. The findings allow initial insights into the perception of automotive UI from different international perspectives. The next consequential step is to provide more finely grained information on individual aspects, such as differential successes and failures during the interaction, to determine the underlying root causes of these differences. One potential reason could be grounded in cultural differences between user groups in international markets. Since culture impacts perception [65–67] and perception in turn affects evaluations, culture should be considered an influential factor in the perception of an automotive interface [6]. Thus, culture seems to be a feasible factor to explain differences in human-machine interactions. Future research should be conducted in diverse cultural contexts that differ in most of Hofstede's six cultural dimensions to ensure that they represent different sets of values and, hence, that they represent different cultures. Ensuring that the user groups are indeed representative of different cultures and thus attributing the observed differences to cultural variances exceeds the scope of the current findings. However, this is also not the aim of the benchmark data presented.

Besides choosing countries for future studies, researchers must define which components of the HMI will be further analyzed. IVISs include navigation, radio, or communication, among others. Mehler and colleagues [45], for example, focused on differences between German and Chinese users in the design of the navigation map. Lindgren and colleagues [16] identified the needs and requirements of the design of advanced driving assistance systems for Swedish and Chinese drivers, and Young and colleagues [22] identified different preferences for IVIS control types and labels (i.e., fan speed control) for Australian and Chinese drivers. Another HMI element that should be addressed in cross-cultural studies is the voice user interface (VUI). Gong and colleagues [54] already stated that Chinese customers would be willing to pay more for certain functions such as a physical voice assistant. It is yet unclear and needs to be investigated whether customers from Western markets desire those VUI features, too. Furthermore, instead of conducting another field study, future studies should instead comprise driving simulators. To better understand the differences between countries in UI evaluation, a more standardized setting should be chosen to control certain variables. Different factors can be included in a between-within study design to analyze specific impacts on the dependent measures, i.e., *market* as a between-subjects factor and, in the case of VUI, *visualization* as a within-subject factor. As outlined above (see Section 4.4), the results suggest that a triangulation of methods is necessary and that it is important to include both self-report as well as observational measures for a comprehensive understanding of human-machine interactions.

5. Conclusions

In conclusion, the results indicate that differences in the evaluation of an automotive UI exist between different user groups since the ratings differ both in self-report and behavioral measures. The preferences of specific user groups shall therefore be taken into consideration when designing interfaces for different international markets. For global automotive industries, the results highlight the challenge of respecting user expectations and needs in diversified markets while still benefiting from global processes. The results suggest that market-specific adaptations of HMI could be promising to enhance preference and interaction performance.

Author Contributions: Conceptualization, D.S., Y.F., F.N. and J.F.K.; methodology, D.S., Y.F., F.N. and J.F.K.; formal analysis, D.S.; data curation; D.S., writing—original draft preparation, D.S.; writing—review and editing, D.S., Y.F., F.N. and J.F.K.; visualization, D.S.; supervision, J.F.K. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not need ethical approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not available due to privacy reasons.

Conflicts of Interest: Authors Denise Sogemeier, Yannick Forster, Frederik Naujoks and Andreas Keinath were employed by the company BMW Group. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Table A1. Descriptive statistics (i.e., adjusted *M*, *SE*) for the interaction performance scores in study 2 for (a) the three UCs and (b) for the two modalities split by market.

(a)				(b)			
Market	Use Case	Adjusted <i>M</i>	<i>SE</i>	Market	Modality	Adjusted <i>M</i>	<i>SE</i>
Germany	Navigation	1.67	0.14	Germany	Touch	1.28	0.08
	Media	1.26	0.10		Remote	1.53	0.13
	Communication	1.28	0.12				
China	Navigation	2.68	0.14	China	Touch	1.42	0.08
	Media	1.27	0.10		Remote	2.18	0.13
	Communication	1.45	0.12				

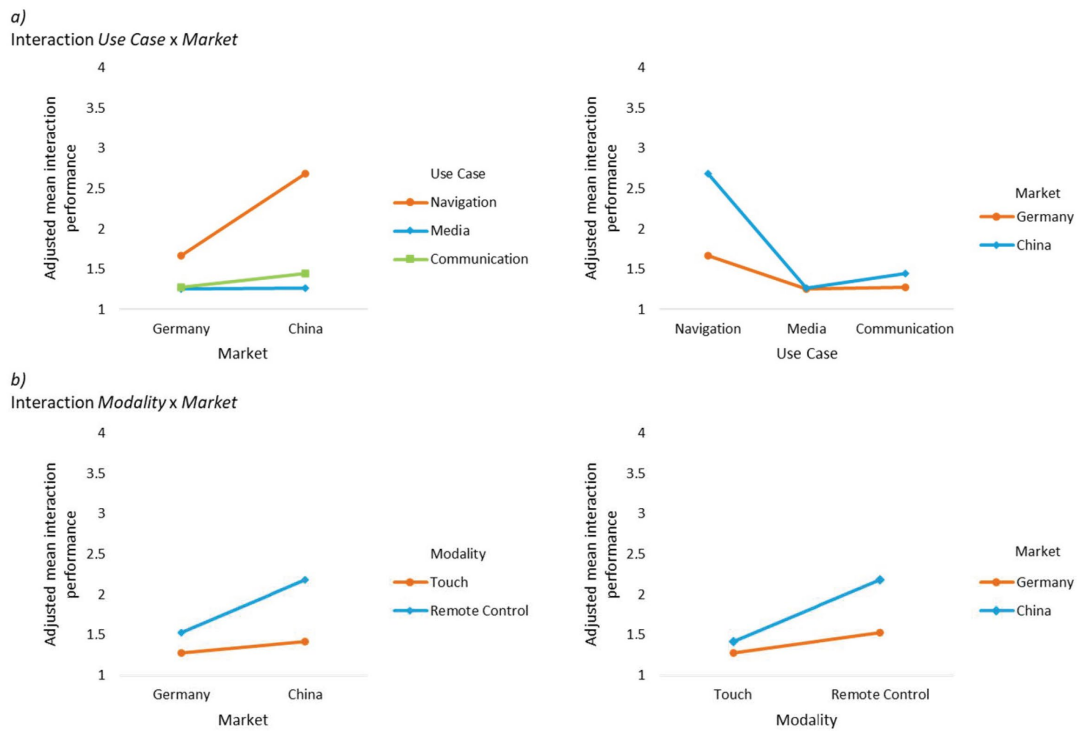


Figure A1. Adjusted mean interaction performance scores to visualize the interaction effect between (a) UC and market and (b) modality and market for study 2.

Appendix B

Table A2. Descriptive statistics (i.e., adjusted *M*, *SE*) for the interaction performance scores in study 3 for the three UCs split by market.

Market	Use Case	Adjusted <i>M</i>	<i>SE</i>
Germany	Navigation	2.27	0.16
	Media	1.23	0.07
	Communication	1.28	0.11
China	Navigation	1.40	0.16
	Media	1.05	0.08
	Communication	1.18	0.12
US	Navigation	1.20	0.16
	Media	1.13	0.07
	Communication	1.36	0.11

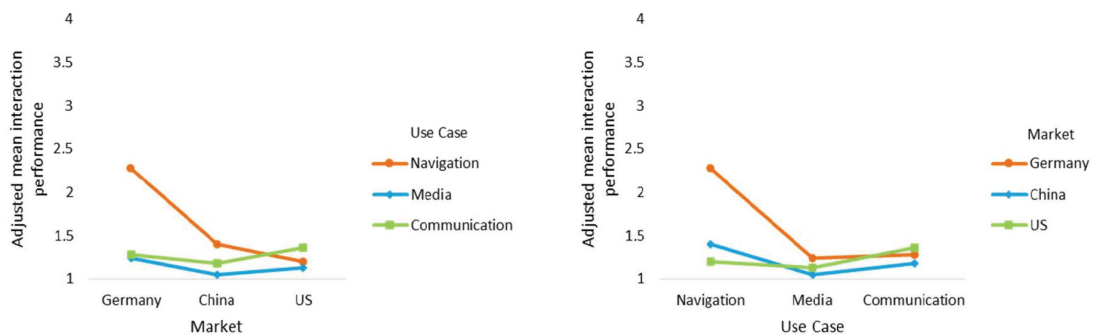


Figure A2. Adjusted mean interaction performance scores to visualize the interaction effect between UC and market for study 3.

Appendix C

Table A3. Inferential statistics (i.e., $df1$, $df2$, F -, p -, and η_p^2 -values) for the mixed between-within ANCOVA for experimenter ratings in study 3 (modality: remote control).

Effect	$df1$	$df2$	F	p	η_p^2
Market	1	79	0.04	0.842	0.001
Age	1	79	5.82	0.018	0.07
UC	1.38	109.30	1.49	0.230	0.02
UC \times Market	2	158	0.02	0.982	0.00
UC \times Age	2	158	0.54	0.585	0.01

Note: significant effects are in bold.

References

1. Watson, G.H. *Strategic Benchmarking: How to Rate Your Company's Performance Against the World's Best*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1993.
2. Sweeney, M.; Maguire, M.; Shackel, B. Evaluating user-computer interaction: A framework. *Int. J. Man-Mach. Stud.* **1993**, *38*, 689–711. [[CrossRef](#)]
3. Kumar, A.; Antony, J.; Dhakar, T.S. Integrating quality function deployment and benchmarking to achieve greater profitability. *BIJ* **2006**, *13*, 290–310. [[CrossRef](#)]
4. Erdil, A.; Erbiyik, H. The Importance of Benchmarking for the Management of the Firm: Evaluating the Relation between Total Quality Management and Benchmarking. *Procedia Comput. Sci.* **2019**, *158*, 705–714. [[CrossRef](#)]
5. Anand, G.; Kodali, R. Benchmarking the benchmarking models. *Benchmarking Int. J.* **2008**, *15*, 257–291. [[CrossRef](#)]
6. Rössger, P. Intercultural HMIs in Automotive: Do We Need Them?—An Analysis. In *HCI International 2021—Late Breaking Papers: Design and User Experience*; Stephanidis, C., Soares, M.M., Rosenzweig, E., Marcus, A., Yamamoto, S., Mori, H., Rau, P.-L.P., Meiselwitz, G., Fang, X., Moallem, A., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 584–596, ISBN 978-3-030-90237-7.
7. Kern, D.; Schmidt, A. Design Space for Driver-based Automotive User Interfaces. In Proceedings of the AutomotiveUI'09: 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Essen, Germany, 21–22 September 2009.
8. Young, K.; Regan, M. Driver Distraction: A Review of the Literature. In *Distracted Driving*; Australasian College of Road Safety: Sydney, NSW, Australia, 2007.
9. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.; Keinath, A. User Education in Automated Driving: Owner's Manual and Interactive Tutorial Support Mental Model Formation and Human-Automation Interaction. *Information* **2019**, *10*, 143. [[CrossRef](#)]
10. *ISO 9241-210:2010*; Prozess zur Gestaltung Gebrauchstauglicher Interaktiver Systeme. Deutsches Institut für Normung: Berlin, Germany, 2011.
11. Forster, Y.; Kraus, J.M.; Feinauer, S.; Baumann, M. Calibration of Trust Expectancies in Conditionally Automated Driving by Brand, Reliability Information and Introductory Videos. In Proceedings of the AutomotiveUI'18: 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Toronto, ON, Canada, 23–25 September 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 118–128, ISBN ISBN 9781450359467.
12. Norman, D.A. Human-Centered Design Considered Harmful. *Interactions* **2005**, *12*, 14–19. [[CrossRef](#)]
13. Hassenzahl, M. The Effect of Perceived Hedonic Quality on Product Appealingness. *Int. J. Hum. Comput. Interact.* **2001**, *13*, 481–499. [[CrossRef](#)]
14. Forster, Y.; Hergeth, S.; Naujoks, F.; Krems, J.F.; Keinath, A. Self-report measures for the assessment of human-machine interfaces in automated driving. *Cogn. Technol. Work* **2020**, *22*, 703–720. [[CrossRef](#)]
15. Pettersson, I.; Lachner, F.; Frison, A.-K.; Riener, A.; Butz, A. A Bermuda Triangle?—A Review of Method Application and Triangulation in User Experience Evaluation. In Proceedings of the CHI'18: CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; Mandryk, R., Hancock, M., Perry, M., Cox, A., Eds.; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–16, ISBN ISBN 9781450356206.
16. Lindgren, A.; Chen, F.; Jordan, P.W.; Zhang, H. Requirements for the Design of Advanced Driver Assistance Systems—The Difference between Swedish and Chinese Drivers. *Int. J. Des.* **2008**, *2*, 41–54.
17. Heimgaertner, R. Towards Cultural Adaptability in Driver Information and -Assistance Systems. In *Usability and Internationalization Part II*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 372–381.
18. Lesch, M.F.; Rau, P.-L.P.; Zhao, Z.; Liu, C. A cross-cultural comparison of perceived hazard in response to warning components and configurations: US vs. China. *Appl. Ergon.* **2009**, *40*, 953–961. [[CrossRef](#)]
19. Khan, T.; Williams, M. A Study of Cultural Influence in Automotive HMI: Measuring Correlation between Culture and HMI Usability. *SAE Int. J. Passeng. Cars-Electron. Electr. Syst.* **2014**, *7*, 430–439. [[CrossRef](#)]

20. Wang, M.; Lyckvi, S.L.; Chen, F. Same, Same but Different: How Design Requirements for an Auditory Advisory Traffic Information System Differ Between Sweden and China. In Proceedings of the AutomotiveUI'16: 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Ann Arbor, MI, USA, 24–26 October 2016; Green, P., Boll, S., Gabbard, J., Osswald, S., Burnett, G., Borojeni, S.S., Löcken, A., Pradhan, A., Eds.; Association for Computing Machinery: New York, NY, USA, 2016; pp. 75–82, ISBN ISBN 9781450345330.
21. Khan, T.; Williams, M. Cross-Cultural Differences in Automotive HMI Design: A Comparative Study Between UK and Indian Users' Design Preferences. *J. Usability Stud.* **2016**, *11*, 45–65.
22. Young, K.L.; Rudin-Brown, C.M.; Lenné, M.G.; Williamson, A.R. The implications of cross-regional differences for the design of In-vehicle Information Systems: A comparison of Australian and Chinese drivers. *Appl. Ergon.* **2012**, *43*, 564–573. [[CrossRef](#)] [[PubMed](#)]
23. Braun, M.; Li, J.; Weber, F.; Pfleging, B.; Butz, A.; Alt, F. What If Your Car Would Care? Exploring Use Cases For Affective Automotive User Interfaces. In Proceedings of the MobileHCI'20: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, Oldenburg, Germany, 5–8 October 2020; Association for Computing Machinery: New York, NY, USA, 2020. ISBN ISBN 9781450375160.
24. Zhu, D.; Wang, D.; Huang, R.; Jing, Y.; Qiao, L.; Liu, W. User Interface (UI) Design and User Experience Questionnaire (UEQ) Evaluation of a To-Do List Mobile Application to Support Day-To-Day Life of Older Adults. *Healthcare* **2022**, *10*, 2068. [[CrossRef](#)] [[PubMed](#)]
25. Niklas, U.; von Behren, S.; Eisenmann, C.; Chlond, B.; Vortisch, P. Premium factor—Analyzing usage of premium cars compared to conventional cars. *Res. Transp. Bus. Manag.* **2019**, *33*, 100456. [[CrossRef](#)]
26. François, M.; Osiurak, F.; Fort, A.; Crave, P.; Navarro, J. Automotive HMI design and participatory user involvement: Review and perspectives. *Ergonomics* **2016**, *60*, 541–552. [[CrossRef](#)] [[PubMed](#)]
27. Brooke, J. SUS—A quick and dirty usability scale. *Usability Eval. Ind.* **1996**, *189*, 4–7.
28. Li, R.; Chen, Y.V.; Sha, C.; Lu, Z. Effects of interface layout on the usability of In-Vehicle Information Systems and driving safety. *Displays* **2017**, *49*, 124–132. [[CrossRef](#)]
29. Forster, Y.; Hergeth, S.; Naujoks, F.; Beggiato, M.; Krems, J.F.; Keinath, A. Learning to use automation: Behavioral changes in interaction with automated driving systems. *Transp. Res. Part F Traff. Psychol. Behav.* **2019**, *62*, 599–614. [[CrossRef](#)]
30. Gao, M.; Kortum, P.; Oswald, F.L. Multi-Language Toolkit for the System Usability Scale. *Int. J. Hum. Comput.* **2020**, *36*, 1883–1901. [[CrossRef](#)]
31. Lewis, J.R. The System Usability Scale: Past, Present, and Future. *Int. J. Hum. Comput.* **2018**, *34*, 577–590. [[CrossRef](#)]
32. Loew, A.; Sogemeier, D.; Kulesa, S.; Forster, Y.; Naujoks, F.; Keinath, A. A Global Questionnaire? An International Comparison of the System Usability Scale in the Context of an Infotainment System. In Proceedings of the International Conference on Applied Human Factors and Ergonomics. International Conference on Applied Human Factors and Ergonomics, New York, NY, USA, 24–28 July 2022; pp. 224–232.
33. Laugwitz, B.; Held, T.; Schrepp, M. Construction and Evaluation of a User Experience Questionnaire. In Proceedings of the HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, Graz, Austria, 20–21 November 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 63–76.
34. Schankin, A.; Budde, M.; Riedel, T.; Beigl, M. Psychometric Properties of the User Experience Questionnaire (UEQ). In Proceedings of the CHI'22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; Barbosa, S., Lampe, C., Appert, C., Shamma, D.A., Drucker, S., Williamson, J., Yatani, K., Eds.; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–11, ISBN ISBN 9781450391573.
35. Greve, G.; Benning-Rohnke, E. (Eds.) *Kundenorientierte Unternehmensführung: Konzept und Anwendung des Net Promoter® Score in der Praxis*, 1st ed.; Gabler: Wiesbaden, Germany, 2010; ISBN 9783834923196.
36. Naujoks, F.; Wiedemann, K.; Schömig, N.; Jarosch, O.; Gold, C. Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX* **2018**, *5*, 579–592. [[CrossRef](#)] [[PubMed](#)]
37. Kenntner-Mabiala, R.; Kausner, Y.; Hoffmann, S.; Volk, M. Driving performance of elderly drivers in comparison to middle-aged drivers during a representative, standardized driving test in real traffic. *Z. Für Verkehrssicherheit* **2016**, *3*, 73.
38. Jarosch, O.; Bengler, K. Rating of Take-Over Performance in Conditionally Automated Driving Using an Expert-Rating System. In *Advances in Human Aspects of Transportation*; Stanton, N., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 283–294, ISBN 978-3-319-93884-4.
39. Forster, Y. Preference versus Performance in Automated Driving: A Challenge for Method Development. Doctoral Thesis, University of Technology Chemnitz, Chemnitz, Germany, 2019.
40. Field, A.P. *Discovering Statistics Using SPSS*, 3rd ed.; SAGE Publications: Los Angeles, CA, USA, 2009; ISBN 9781847879066.
41. Armstrong, R.A. When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* **2014**, *34*, 502–508. [[CrossRef](#)] [[PubMed](#)]
42. Reichheld, F.F. The one number you need to grow. *Harv. Bus. Rev.* **2003**, *81*, 46–55. [[PubMed](#)]
43. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; L. Erlbaum Associates: Hillsdale, NJ, USA, 1988.

44. Heimgärtner, R. ISO 9241-210 and Culture?—The Impact of Culture on the Standard Usability Engineering Process. In *Design, User Experience, and Usability. User Experience Design Practice*; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Kobsa, A., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Rangan, C.P., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 39–48, ISBN 978-3-319-07637-9.
45. Mehler, J.; Guo, Z.; Zhang, A.; Rau, P.-L.P. Quick Buttons on Map-Based Human Machine Interface in Vehicles is Better or Not: A Cross-Cultural Comparative Study Between Chinese and Germans. In *Culture and Computing. Design Thinking and Cultural Computing*; Rauterberg, M., Ed.; Springer International Publishing: Cham, Switzerland, 2021; pp. 432–449, ISBN 978-3-030-77430-1.
46. Hofstede, G. Dimensionalizing Cultures: The Hofstede Model in Context. *Psychol. Cult.* **2011**, *2*, 8. [[CrossRef](#)]
47. Hofstede, G.J. Sixth Dimension Synthetic Culture Profiles. 2010. Available online: <https://geerthofstede.com/wp-content/uploads/2016/08/sixth-dimension-synthetic-culture-profiles.doc> (accessed on 11 May 2022).
48. Trompenaars, F.; Hampden-Turner, C. *Riding the Waves of Cultures: Understanding Cultural Diversity in Business*; Nicholas Brealey: London, UK, 2011.
49. Chang, S.; Kim, C.-Y.; Cho, Y.S. Sequential effects in preference decision: Prior preference assimilates current preference. *PLoS ONE* **2017**, *12*, e0182442. [[CrossRef](#)] [[PubMed](#)]
50. De Souza, T.R.C.B.; Bernardes, J.L. The Influences of Culture on User Experience. In *Cross-Cultural Design*; Rau, P.-L.P., Ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 43–52, ISBN 978-3-319-40092-1.
51. Schrepp, M.; Hinderks, A.; Thomaschewski, J. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Kobsa, A., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Rangan, C.P., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 383–392, ISBN 978-3-319-07667-6.
52. Sogemeier, D.; Forster, Y.; Naujoks, F.; Krems, J.F.; Keinath, A. How to Map Cultural Dimensions to Usability Criteria: Implications for the Design of an Automotive Human-Machine Interface. In Proceedings of the AutomotiveUI'22: 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Seoul, Republic of Korea, 17–20 September 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 123–126, ISBN 9781450394284.
53. Hofstede, G. Culture and Organizations. *Int. Stud. Manag. Organ.* **1980**, *10*, 15–41. [[CrossRef](#)]
54. Gong, Z.; Ma, J.; Zhang, Q.; Ding, Y.; Liu, L. Automotive HMI Guidelines For China Based On Culture Dimensions Interpretation. In *HCI International 2020—Late Breaking Papers: Digital Human Modeling and Ergonomics, Mobility and Intelligent Environments Proceedings of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020*; LNCS; Springer: Cham, Switzerland, 2020; pp. 96–110.
55. Hulstsch, D.F.; MacDonald, S.W.; Dixon, R.A. Variability in Reaction Time Performance of Younger and Older Adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **2002**, *57B*, 101–115. [[CrossRef](#)]
56. Sharit, J.; Czaja, S.J. Ageing, computer-based task performance, and stress: Issues and challenges. *Ergonomics* **1994**, *37*, 559–577. [[CrossRef](#)]
57. Dingus, T.A.; Hulse, M.C.; Mollenhauer, M.A.; Fleischman, R.N.; McGehee, D.V.; Manakkal, N. Effects of Age, System Experience, and Navigation Technique on Driving with an Advanced Traveler Information System. *Hum. Factors* **1997**, *39*, 177–199. [[CrossRef](#)] [[PubMed](#)]
58. Lerner, N.; Singer, J.; Huey, R. *Driver Strategies for Engaging in Distracting Tasks Using In-Vehicle Technologies HS DOT 810 919*; U.S. Department of Transportation: Washington, DC, USA, 2008.
59. Totzke, I. Einfluss des Lernprozesses auf den Umgang mit Menügesteuerten Fahrerinformationssystemen. Doctoral Thesis, Julius-Maximilians-Universität Würzburg, Würzburg, Germany, 2013.
60. Roberts, M.J.; Gray, H.; Lesnik, J. Preference versus performance: Investigating the dissociation between objective measures and subjective ratings of usability for schematic metro maps and intuitive theories of design. *Int. J. Hum. Comput.* **2017**, *98*, 109–128. [[CrossRef](#)]
61. Knapp, B. Mental Models of Chinese and German Users and Their Implications for MMI: Experiences from the Case Study Navigation System. In *Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 882–890.
62. Law, E.L.-C.; van Schaik, P.; Roto, V. Attitudes towards User Experience (UX) Measurement. *Int. J. Hum. Comput.* **2014**, *72*, 526–541. [[CrossRef](#)]
63. Forster, Y.; Naujoks, F.; Neukum, A. Increasing anthropomorphism and trust in automated driving functions by adding speech output. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 365–372, ISBN ISBN 978-1-5090-4804-5.
64. Bortz, J.; Döring, N. *Forschungsmethoden und Evaluation*, 4th ed.; Springer: Berlin/Heidelberg, Germany, 2006; ISBN 978-3-540-33305-0.
65. Boduroglu, A.; Shah, P.; Nisbett, R.E. Cultural Differences in Allocation of Attention in Visual Information Processing. *J. Cross-Cult. Psychol.* **2009**, *40*, 349–360. [[CrossRef](#)] [[PubMed](#)]

66. Hadders-Algra, M. Human face and gaze perception is highly context specific and involves bottom-up and top-down neural processing. *Neurosci. Biobehav. Rev.* **2022**, *132*, 304–323. [[CrossRef](#)]
67. Nisbett, R.E.; Miyamoto, Y. The influence of culture: Holistic versus analytic perception. *Trends Cogn. Sci.* **2005**, *9*, 467–473. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.