UX Metrics: Deriving Country-Specific Usage Patterns of a Website Plug-In from Web Analytics

Florian Lachner, Florian Fincke, Andreas Butz

UX Metrics: Deriving Country-Specific Usage Patterns of a Website Plug-in From Web Analytics

Florian Lachner, Florian Fincke, and Andreas Butz

LMU Munich, Human-Computer Interaction Group
Amalienstr. 17, 80333 Munich, Germany
{florian.lachner,butz}@ifi.lmu.de
florian.fincke@campus.lmu.de

Abstract. Metrics for User Experience (UX) often involve traditional usability aspects, such as task success, but also mental aspects, such as interpretation and meaning. The actual experience of a user also highly depends on personal characteristics, such as the social and cultural background. In this paper, we investigate the relation between users' country of origin and their interaction patterns with an e-commerce website plug-in. We used a quantitative web analytics approach based on six UX-related metrics to evaluate the applicability of a quantitative UX evaluation approach in an international context. In a 34 day study we analyzed the usage patterns of 5.843 French, 2.760 German, and 5.548 Italian website visitors and found that they show significantly different patterns. This indicates that website metrics are a suitable means for cost-effective UX analysis on a large scale, which can provide valuable starting points for a further in-depth analysis.

Keywords: User experience; cross-cultural design; user tracking; data logging; interfaces; globalization; localization

1 Introduction

The theory of User Experience (UX) goes back to the consideration of pleasure and emotions as part of a product's characteristics. Early approaches emerged from a user-centered design perspective, and the awareness of human factor professionals that user satisfaction is insufficiently considered in the concept of usability [26]. The consideration of pleasure and emotions was further increased by the focus on the interplay between affect and cognition. Due to this enhanced view on product design and development, aesthetics, pleasure, and usability became a balanced triad in the HCI community [40].

Nowadays, the primary goal of UX designers and engineers often is to create a pleasurable interaction between the user and the product that goes beyond traditional usability considerations [19]. It also has become common ground in the HCI community that experiences are subjective in nature and highly dependent on the usage context [24,32]. Hence, a user's experiences can be shaped and influenced based on his or her individual preferences (regarding aesthetics or ergonomics), mood, prior interactions, product brand, age, gender, and culture [7,12,16,29,30,34,42,51,52]. The cultural aspect becomes particularly interesting for global businesses, whose products or services can be accessed, evaluated, and purchased from all over the world [17,37,46,61].

In order to ensure the intended quality of UX, measurement tools and methods represent a crucial resource in UX design and research processes. However, there is still an ongoing debate about the applicability and effectiveness of qualitative and quantitative approaches for UX measurement [6,33,57]. Furthermore, researchers and designers have to balance information value, cost efficiency, and expenditure of time when gathering attitudinal (e.g., through lab studies or surveys) or behavioral data (e.g., through data logging or time measurement) [50,55,56].

In this paper, we analyze the relationship between the country of origin and the usage behavior of users of a website plug-in (see Figure 1). We base our analysis on quantitative behavioral data, gathered through user tracking, to draw a conclusion on the applicability of web analytics metrics. Our dataset stems from a data logging study of a plug-in that was implemented in an e-commerce website plug-in.

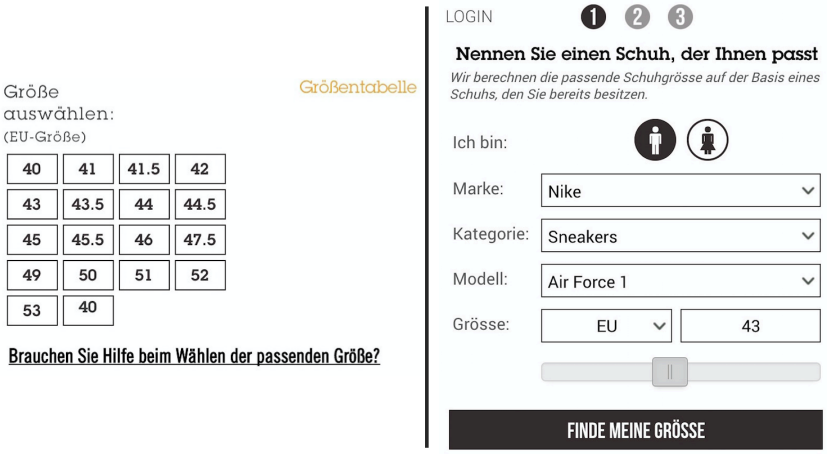

Fig. 1. Website plug-in (right) for shoe size recommendations and the link to it in the German online store (left).

Over the course of 34 consecutive days we tracked the behavior (i.e., plug-in interactions) of users located in France, Germany, and Italy based on six UX metrics, which we derived and adapted from the HEART framework of Rodden et al. [50]. Our study was motivated by the following research question:

Which differences in the user experience of a website (plug-in) can we identify between French, German, and Italian users simply through web analytics metrics?

Consequently, the contribution of this paper is twofold: First, based on the analysis of country-specific differences we identify associated relationships and hence suitable levers to efficiently target further qualitative in-depth analyses. Second, we adapt the quantitative UX framework of Rodden et al. [50] to our specific use case (i.e., e-commerce

website plug-in) in order to examine the applicability of UX metrics that build upon large-scale website tracking data. Ultimately, we draw a conclusion how such a quantitative approach can support designers in saving time and money for cross-cultural UX evaluation and potentially localized interface adaptions. For our analysis, we, therefore, exclude a supplemental investigation of further factors, such as gender differences or device type. Our underlying goal is to foster an ongoing debate about cross-cultural UX design and about an appropriate balance of qualitative and quantitative UX measurement.

2 Background and Related Work

Despite the general agreement on its importance for human-centered design, researchers and practitioners still struggle to narrow down the broad field of UX to one unified definition [20]. The lack of a common definition of UX entails a large variety of research directions in the field of HCI, with foci ranging from usability to psychological needs and emotions [6,32]. To locate our work in this ongoing discourse, the following sections illustrate the basic scope of (our understanding of) UX, some key aspects of UX measurement, as well as related work in the field of cross-cultural UX design.

2.1 The Scope of User Experience

The main difference between usability and UX is that UX researchers and designers can not merely focus on a product's characteristics (i.e., functionality, purpose, etc.) but also have to consider the user's needs and motivation as well as the context of use (i.e., the environment) [9,13,19,24,38]. Consequently, experiences do not only result from interacting with a product but also from a user's expectations, others' opinions, or from experiences with related technologies before the actual interaction. At the same time, experiences and associated feelings merely evolve over time through reflection on previous interactions, advertisements, and again through others' opinions [24,27,35,52].

The scope of UX becomes even more complex for globally acting businesses: First, the concept of UX is differently understood between academia and industry as well as between different countries [31,32]. Second, cultural differences in language, values, or needs raise various questions regarding the suitability of globally optimal or locally adapted designs of products and services [4,37,59].

In this paper, our goal is to analyze cultural differences in UX design. For this purpose, we simplify the origin of cultural differences to individual preferences caused by one's country of origin. Thus, we do not focus on further cultural allocations, such as age group or social background. Furthermore, UX in our context shall include both usability aspects as well as mental aspects, such as the interpretation of an e-commerce website plug-in. In order to answer our research question, we will, therefore, derive suitable web analytics metrics, which we call UX metrics.

2.2 Cross-Cultural Differences in UX Design

The need for cross-cultural considerations in interface design emerged more than two centuries ago, shortly after designers started to put an emphasis on the usability aspects of

their designs (see [39]). Initial discussions mainly focused on the use of colors, language, as well as icons and symbols [5,53]. However, since then usability theories and measures in the HCI community rather marginally focused on cultural design preferences [14]. Nevertheless, with the further increasing interest in experiences of product interactions, researchers in the HCI community once again started to raise questions about cross-cultural design preferences (see [8,21,47]). In fact, various studies have already been able to identify cultural differences in UX design in different use cases.

Athinen et al. [1], for example, investigated culturally sensitive design for a mobile wellness application. In their study, they interviewed 16 people (8 from Finland and 8 from India) to identify similarities and differences in the understanding of wellness and its consequences for the design of a mobile application. They found that Finns and Indians have a different understanding of goal setting, which is an important aspect for the associated mobile application. Similarly, Walsh and Vaino [60] argue for cross-cultural UX considerations for mHealth applications, while Al-Shamaileh and Sutcliffe [2] demonstrate varying preferences in the design of health-related websites in the UK and Jordan.

Furthermore, Frandsen-Thorlacius et al. [14] were able to detect differences in the understanding of the concept of usability for Danish and Chinese users. Using a questionnaire survey, the authors were able to derive that Chinese users preferentially value visual appearance, satisfaction, and fun, whereas Danish users rather focus on effectiveness, lack of frustration, and efficiency. Reinecke and Gajos [48] were, likewise, able to analyze visual preferences of websites based on a comprehensive study of 2.4 million ratings from almost 40 thousand participants.

However, cultural differences are not limited to the evaluation of products and services. Lallemand et al. [31] point out discrepancies in the understanding of the concept of UX based on a survey amongst 758 researchers and practitioners from 35 nationalities. Gerea and Herskovic [15] additionally expand this study to Latin America. Nowadays, researchers want to further link cultural studies and product design, particularly through the integration of Hofstede's (see [22]) cultural dimensions in HCI [43,36,37,46,58].

2.3 Qualitative vs. Quantitative UX Measurement

Because experiences are such a complex phenomenon, UX researchers and practitioners utilize a whole set of measurement approaches to anticipate, test, and improve a product's UX. However, there is no common agreement whether qualitative or quantitative approaches should be favored [6,33,49]. On the one hand, qualitative approaches (gathered through, e.g., interviews) provide rich and detailed insights for in-depth analysis [54], on the other hand, quantitative approaches (gathered through, e.g., questionnaires) can reduce costs and time effort [23,57].

Apart from this, UX measurement methods are primarily based on attitudinal data (i.e., data related to a user's feelings and emotions) [31,50]. In contrast, the HEART framework [50] represents a first step towards the integration of behavioral data (i.e., actual activities of users - traditionally used in usability testing, see [3,10,25,41]), in UX measurement. The framework includes five metrics, focusing on both usability and UX-related aspects [50]:

- *Happiness:* referring to, e.g., satisfaction and ease of use.
- *Engagement:* describing the user's level of involvement.
- *Adoption:* addressing customer acquisition.
- *Retention:* analyzing recurring users.
- *Task success:* covering traditional usability aspects.

The framework does not aim to describe UX as a whole but to strategically direct UX measurement processes based on large-scale data, particularly when working in teams. Therefore, one has to define a suitable measurement goal and approach per metric (e.g., the number of visits per week for *Engagement*, the error rate for *Task success*) depending on the respective product or service.

We understand their approach as an initial step towards including behavioral data from usability testing in UX measurement. Therefore, we aim to evaluate its applicability for our use case, i.e., the analysis of UX-related, country-specific usage patterns of French, German, and Italian users from web analytics. However, to ensure a suitable implementation of UX metrics in our collaboration partner's development process, we slightly customized our UX metrics based on the HEART framework.

3 Methodology and User Study

In order to examine the applicability of UX-oriented web analytics metrics for identifying country-specific user behaviors, we partnered with a company that provides a customizable website plug-in for online shoe stores. The plug-in allows customers to identify their correct shoe size based on the comparison with the size of another model.

3.1 Setting and Procedure

For our study, we tracked the plug-in interactions of a globally acting online shoe store. The analyzed plug-in (see Figure 1 right) is integrated in the store's website and accessible through a link below the actual selection of the shoe size (see Figure 1 left). The overall goal of the plug-in is that customers can enter information about a shoe that they already own in order to identify the correct size of the shoe they want to buy. To ensure a problem-free implementation in different countries, the plug-in was translated by professional translators for all countries.

Once a customer clicks on the link, the plug-in opens and asks for the customer's gender as well as the brand, category, model, and size of a comparative shoe (i.e., plug-in steps one to five). This information is used to identify the correct size for the customer depending on the shape and differences in size of the desired shoe. The comparative data is taken from our partner's internal database. As a sixth plug-in step, users can request (i.e., click) a shoe size recommendation. After receiving all the information, the recommended size is stored for 90 days and additionally displayed within the link's text label once a customer accesses the online store again. Thus, it is not necessary to open and use the plug-in repeatedly.

For post-hoc analysis, all tracked data points (plug-in openings, plug-in interactions, recommendation requests, and adding products to the website shopping cart) were

anonymized and securely stored at our partner's server infrastructure for long-term evaluations through client-based tracking. Client-based tracking (i.e., Javascript-based for plug-in interactions and cookie-based tracking for long-term analysis of recurring users) was pursued to minimize data traffic in order to ensure a smooth and pleasant plug-in implementation. Shoe recommendations were tracked through server-based tracking. The country of origin was identified by the client's IP address.

3.2 Study Data and Analysis

Observations. We ran our study for 34 consecutive days. During this time, no special offer or promotion was announced at the client's online store in order to ensure the comparability of our analysis. Over the course of our study people from 200 countries visited the client's website, whereof people from 121 countries accessed the plug-in. For our investigation we focused on France, Germany, and Italy (277,551, 141,897, and 172,887 website loadings leading to 5843, 2760, and 5548 plug-in openings, respectively). Overall, about one third (31,4% in France, 30,4% in Germany, and 37,2% in Italy) of all website visitors per country accessed the website on a mobile device, two thirds (68,6% in France, 69,6% in Germany, and 62,8% in Italy) on a desktop device.

UX metrics. Our quantitative analysis of the plug-in interactions was based on six metrics (see Table 1) that we derived and adapted from the HEART framework [50]. Our metrics were consciously labeled with a distinguishing term in order to highlight the objective of each metric. Furthermore, the particular term allowed our collaboration partner to align strategic initiatives and development efforts.

Table 1. UX metrics used for the analysis of plug-in interactions.

UX Metric	**Definition and Objective**	*see HEART [50]*
Adoption	No. of openings (link clicks) to measure user acquisition	*Adoption*
Complexity	Time per data input to analyze complexity per plug-in step	*Engagement*
Task Success	No. of total recommendations to track plug-in effectiveness	*Task Success*
Continuity	No. of successful inputs per step to retrace plug-in continuity	*Task Success*
Trust	No. of recommended orders to derive trust in suggestions	*Happiness*
Mastery	No. of suggested orders without plug-in opening (recurring users) to derive long-term trust	*Retention*

First of all, we tracked the user **Adoption**, i.e., the number of users that click on the link to the plug-in as well as the **Complexity** of the plug-in (based on the process time per plug-in step). In order to analyze the effectiveness of the plug-in, we defined the two metrics, **Task Success** (overall number of final recommendations) and **Continuity** (successful completions per plug-in step). These four metrics describe usability aspects of the plug-in.

For the interaction with the online shoe store plug-in, we wanted the associated UX to be a pleasant interaction with the service that results in a trustworthy shoe size

recommendation. The goal of the plug-in recommendation, therefore, is that customers identify the correct size of a shoe and trust the plug-in even when the recommendation differs from the size of the comparative shoe. An additional feature of the plug-in is that the recommended shoe size is stored and shown in the plug-in link when users complete all plug-in steps and access the website again within 90 days (see Figure 2).

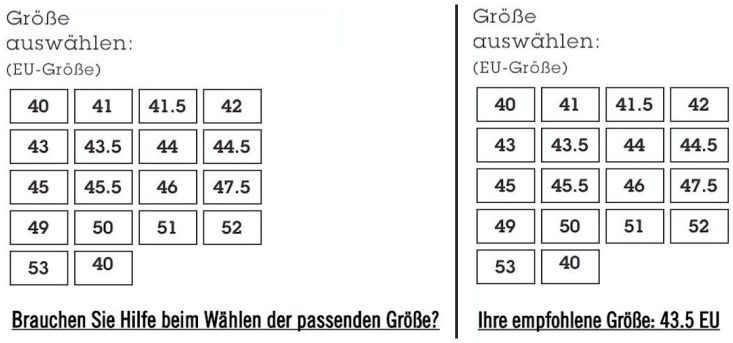

Fig. 2. Link (in the German online store) to the plug-in without recommendation (left) and with recommendation for recurring users (right).

Against this background, we defined the metric **Trust** to understand if users rely on the shoe size recommendation of the plug-in (i.e., put the recommended shoe size into the website's shopping cart). We, therefore, only considered users who ordered a recommended shoe size that differed from the initially entered size of the comparative shoe and excluded users whose recommended size corresponded to the size of the selected comparative shoe. Thus, we could evaluate if users clearly relied on the plug-in's recommendation. We adapted the metric happiness from the HEART framework to our use case as it was not desired to establish a direct communication with the user. All users who successfully clicked through all steps received a recommendation whereas we defined a pleased user as a user that relied on the recommended size for his/her final order. In order to draw conclusions on the long-term experience with the recommendation service, we defined the metric **Mastery**. This metric refers to the number of orders (of recommended shoe sizes) from recurring users that did not open the plug-in again but relied on the suggestion of a suitable size based on their previously entered information. The information was stored in a client-side cookie for 90 days as described before. For this purpose, the recommended shoe size was shown in the link's text label. Once again, we only considered orders that included differing shoe sizes.

All in all, we see these metrics as suitable measuring points for the UX evaluation of equivalent recommendation plug-ins (with the objective to minimize recurring interactions) in an e-commerce context. In further use cases, researchers and designers will have to question their generalization and adapt the metrics accordingly (e.g., when a repetitious interaction is aspired).

Data analysis. We conducted a statistical analysis (using SPSS version 20.0) to identify varying usage behaviors between French, German, and Italian users. We used the Chi-Square Test in order to analyze the association between the country and the UX metrics of *Adoption*, *Task Success*, *Continuity*, *Trust*, and *Mastery*. In order to evaluate the UX metric *Complexity* we used two-way ANOVA and post-hoc Sidak as well as an ANOVA test. We excluded outliers in the process times for the analysis of the metric *Complexity* according to Grubbs [18]. An identified outlier was also excluded from the analysis of previous plug-in steps to ensure consistency within our results. For all analyses we defined a significance level of 5%.

4 Results

The analysis of our data set using the previously defined UX metrics yielded a number of differences in the usage behaviours of the website plug-in between French, German, and Italian users. Thus, we were able to derive significant differences in the adoption rate, dropout rate per plug-in step, the temporal usage patterns, and the reliance on recommendations as described below.

4.1 Country-specific Adoption, Dropout, and Recommendation Rate

First of all, it should be noted that we found a relationship between the country of origin and the *Adoption* rate (see Table 2), i.e., number of plug-in openings to measure user acquisition ($\chi^2(2)=714.327$, p=.000): 2.11% for French users (277,551 website loadings, 5,843 openings), 1.95% for German users (141,897 website loadings, 2,760 openings), and 3.21% for Italian users (172,887 website loadings, 5,548 openings).

The analysis of *Continuity* (i.e., number of successful inputs per plug-in step to retrace usage continuity) provided insights in the relationship between country of origin and successful completions per plug-in step. We found a relationship for the plug-in steps where users had to select their gender ($\chi^2(4)=28.267$, p=.000), the brand of a comparative shoe ($\chi^2(4)=10.166$, p=.038), an associated model ($\chi^2(4)=22.019$, p=.000), and click to receive a shoe size recommendation ($\chi^2(2)=6.781$, p=.034), as summarized in Table 2.

Except for the last step, where users had to click to receive a recommendation, we included users who successfully completed the respective step (success), closed the plug-in or browser (failure), and users who went back to the respective plug-in step after having already moved on to further plug-in steps (detour) in the analysis of the usage *Continuity*. Thus, we were able to derive usage patterns per plug-in step: Generally, in the first plug-in step (i.e., selection of the gender) users from all three countries showed the highest dropout rate (including only successful and failed completions): 22.51% for France (1315 failed users), 24.82% for Germany (685 failed users), and 26.71% for Italy (1482 failed users). In addition, most users who went back to a previous plug-in step chose to start from the beginning of entering the comparative data, more precisely by selecting the brand (the second plug-in step) of a comparative shoe (see Table 2).

Table 2. Chi-Square results (χ^2) based on the UX metrics Adoption, Continuity, Task Success, Trust, and Mastery for French (FRA), German (GER), and Italian (ITA) users.

UX metric	Plug-In Step		Country of Origin				χ^2	Cramer's V
			FRA	*GER*	*ITA*	*Total*		
Adoption		*Success*	5843	2760	5548	14.151	714.327	.035
		Failure	271.708	139.137	167.339	578.184		
		Total	277.551	141.897	172.887	592.335	p=.000	
Continuity	Gender	*Success*	4528	2075	4066	10,669	28.267	.032
		Failure	1315	685	1482	3482		
		Detour	14	8	19	41		
		Total	5857	2768	5567	14,192	p=.000	
	Brand	*Success*	4105	1917	3733	9755	10.166	.022
		Failure	183	72	134	389		
		Detour	320	120	308	748		
		Total	4608	2109	4175	10.892	p=.038	
	Category	*Success*	3908	1797	3540	9245	4.218	n.s.
		Failure	312	155	283	750		
		Detour	178	87	196	461		
		Total	4398	2039	4019	10.456	p=.377	
	Model	*Success*	3354	1516	2906	7776	22.019	.034
		Failure	664	325	728	1717		
		Detour	27	18	44	89		
		Total	4045	1859	3678	9582	p=.000	
	Size	*Success*	3159	1417	2698	7274	8.256	n.s.
		Failure	213	108	228	549		
		Detour	6	4	12	22		
		Total	3378	1529	2938	7845	p=.083	
	Rec.	*Success*	3038	1350	2560	6948	6.781	.031
		Failure	125	70	145	340		
		Total	3163	1420	2705	7288	p=.034	
Task Success		*Openings*	5843	2760	5548	14.151	13.332	.025
		Rec.	3038	1350	2560	6948		
		Total	8881	4110	8108	21.099	p=.001	
Trust		*Yes*	10	3	9	22	21.232	.193
		No	381	113	53	547		
		Total	391	116	62	569	p=.000	
Mastery		*Yes*	158	17	103	278	42.130	.136
		No	1094	421	500	2015		
		Total	1252	438	603	2293	p=.000	

In addition, we were able to identify a relationship of *Task Success* (i.e., number of total recommendations to understand plug-in effectiveness) and country of origin ($\chi^2(2)=13.332$, p=.001). Users from France showed the highest rate of successful recommendations (52% out of 5843 plug-in openings), followed by Germany (49% out of 2760 plug-in openings), and Italy (46% out of 5548 plug-in openings).

4.2 Divergent Temporal Usage Patterns

The goal of the metric *Complexity* was to identify temporal differences along the process steps in order to diagnose key hurdles of the plug-in. We used the z-score transformation to make the data normal before conducting the (two-way) ANOVA and post-hoc Sidak test, as our dataset (process time per plug-in step) did not represent a normal distribution according to the Kolmogorov-Smirnov test. We used post-hoc Sidak test as all users interacted with the plug-in independently. We then used the two-way ANOVA and a post-hoc Sidak test to analyze the effect of the country of origin on the time spent on each step along the plug-in process. Thus, we found out that there was an effect between country of origin and the process time per plug-in step ($F(10,2)=10.427$, p=.000, $\eta^2=.011$). In our study, Italian users significantly differed in their temporal usage patterns along all plug-in steps from French users (p=.000) as well as from German users (p=.022). French and Germany did not differ significantly (p>.050) (see Figure 3).

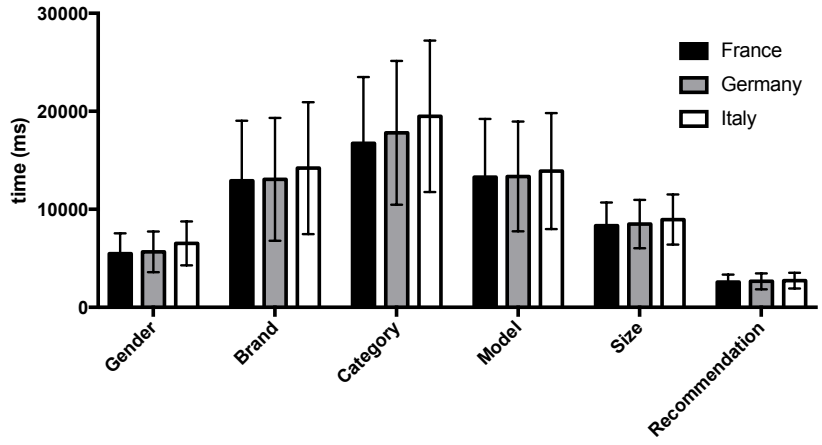

Fig. 3. Average process times for each plug-in step per country including the standard deviation.

Next, we conducted an ANOVA and post-hoc Sidak test to investigate whether the process times significantly vary per plug-in step. Thus, we found out that there is an effect of country of origin for the first plug-in step to select the gender ($F(12,2)=10.774$, p=.000, $\eta^2=.012$) as well as the third plug-in step to select a category of a comparative shoe ($F(12,2)=6.342$, p=.002, $\eta^2=.007$). For the other plug-in steps (i.e., brand, model,

size, and recommendation) we could not identify significant differences (p>.050). More precisely, for the first step (i.e., gender) the process time (i.e., the mean) of Italian users differed from French users (p=.000) as well as from German users (p=.008). Furthermore, the mean of the process time of Italian users to select a category varied from the process time of French users (p=0.001). On average, Italian users needed more time for each plug-in step.

4.3 Varying Reliance on Recommendations

Based on the two metrics *Trust* (number of recommended orders) and *Mastery* (number of recommended orders of recurring users without opening the plug-in) we analyzed the usage behaviors of plug-in users directly related to the recommendation service. The objective of these metrics is to understand whether the country of origin is related to the reliance of users on the shoe size recommendation as well as with the understanding of recurring users (who already successfully clicked through the whole plug-in process and should understand that their suitable size is directly represented in the plug-in opening link) that they do not have to open the plug-in again.

We found out that there is a relationship between country of origin and the *Trust* in the recommendation of the plug-in ($\chi^2(2)=13.983$, p=.001). Furthermore, the country of origin is related to the understanding of the link's text label recommendation (*Mastery*) for recurring users ($\chi^2(2)=42.130$, p=.000).

In our study, French and German users showed a comparable trust rate (i.e., number of users who ordered a differing shoe size based on the recommendation and excluding users whose initially entered shoe size equalled the recommended size hence no conclusion on the user's trust can be drawn) of 2.56% (FR: 10 out of 391, 114 additional users excluded) and 2.59% (GER: 3 out of 116, 60 additional users excluded). However, from 62 Italian users that got a differing recommendation, 9 users (14.52% excluding 27 additional users) relied on the plug-in and added a differing shoe size into the website's shopping cart (see Figure 4).

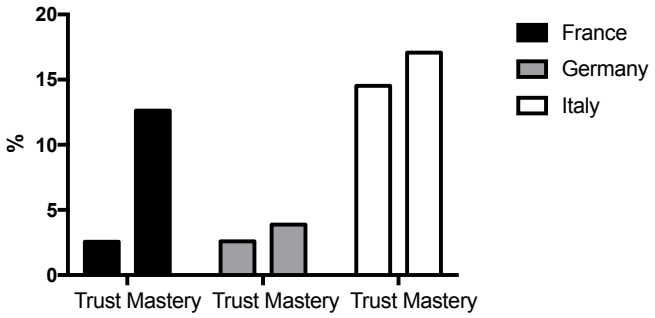

Fig. 4. Percentage of users per country who relied on the plug-in suggestion (Trust) or the suggestion in the link's text label (Mastery) when the recommended size differed from the comparative shoe size.

With regards to recurring users relying on the recommendation of the plug-in link, the number of French users (adding a differing shoe size into the website's shopping cart without opening the plug-in and once again excluding users for whom no conclusion can be drawn as the recommended size equalled the initially entered size) increased to 12.62% (158 out of 1252, excluding 486 additional users) and the number of reliant Italian users increased to 17.08% (103 out of 603, excluding 133 additional users). German users, however, remained at a rather low rate of 3.88% (17 out of 438, excluding 250 additional users) (see Figure 4).

4.4 Summary and Interpretation

The analysis of UX metrics allowed us to understand country-specific usage patterns of French, German, and Italian users. Users from all three countries showed distinct adoption and dropout rates as well as, in particular, significant associations with the plug-in steps gender, brand, model, and recommendation. In addition, we identified significant differences in the overall task success rates. Furthermore, the country of origin is related to the temporal usage patterns along the plug-in steps, with Italian users being the slowest.

Finally, the analysis of the UX-focused metrics *Trust* and *Mastery* showed lower rates for French and German users compared to users from Italy. However, recurring users from France strongly increased their long-term trust rate (i.e., *Mastery*) due to the suggestion in the plug-in link's text label. The described UX metrics helped our project partner to efficiently focus on selected plug-in steps as the analysis highlighted country-specific relationships with a low effect size that are worth paying attention (Cramer's V between .10 and .20 and η^2=.01) compared to country-specific relationships with a marginal effect size (Cramer's V between .00 and .10 and η^2<.01) according to Rea & Parker [45] and Cohen [11]. The localization of all plug-in steps will increase development time and costs. Through the focus on selected and significant plug-in steps with at least low effect sizes, our project partner was able to allocate research and development resources more efficiently.

In order to identify localized interfaces for different countries, designers and researchers need to analyze suitable aspects in further in-depth studies. First, the interface of the website plug-in can be localized and evaluated recurringly for each country to minimize the dropout rate for the critical plug-in steps. One might, for example, prefer text-based icons for the selection of the gender. Second, with regards to the differing process times the plug-in design can be complemented with additional information in order to balance process times per step, dropout rate, and backward steps. Third, it is important to investigate the differences in the *Trust* and *Mastery* rate. German users, for example, might not want to receive suggestions within the link's text label but prefer to receive an individual recommendation each time. Thus, the overall plug-in and link design should be rearranged. Therefore, further qualitative in-depth investigations in the future will allow us to clarify our interpretations.

5 Conclusion, Limitation, and Future Work

In this paper, we demonstrated the applicability of web analytics metrics to analyze differences in the usage behavior and UX of an e-commerce website plug-in between French, German, and Italian users. We were able to identify significant relationships between the country of origin and the adoption rate as well as dropout rate of several plug-in steps. In addition, users from France, Germany, and Italy showed different temporal usage patterns as well as trust in the plug-in's recommendation. Although our work focused on the analysis of an e-commerce plug-in, further country-specific usage patterns have already been identified for Q&A websites such as Yahoo Answers (see [28]) as well as StackOverflow and Superuser (see [44]).

However, narrowing down the complex scope of UX to a selection of six customized website analytics metrics based on client-side user tracking can only be a first step. Overall, it will be necessary to further investigate and analyze the applicability of user tracking for UX measurement due to its quantitatively descriptive nature (see [24,32,41]. Inspired by traditional usability approaches (i.e., logging data) we see our work as a starting point to efficiently guide in-depth UX analyses, complementary to qualitative evaluations with a focus on attitudinal data. Additionally, client-based tracking might not holistically track all website visitors due to, e.g., blocked website cookies. It is, by nature, not possible to track how many website visitors block cookies. We, therefore, limited our analysis to recurring users of plug-in interactions and not website visits. Furthermore, the collaboration with our industry partner did not allow any modification of the original website. Consequently, it was not possible to add a registration process to track the user behavior across different devices.

Based on our research, future studies should add further metrics and qualitative in-depth analysis of country-specific usage patterns, test our findings through locally adapted user interface studies, and investigate the impact of server-based tracking on both the users' UX and the validity of web analytics metrics. Furthermore, the investigation of user-level data (i.e., the consolidated usage data of individual users) might allow conclusions about more detailed user behaviors. Ultimately, to set up a holistic UX-focused user tracking process, it is necessary to compare the effect of cross-country differences with and in contrast to further aspects, such as gender and device type.

We conclude that user tracking can be an efficient way to identify UX-related levers for culturally sensitive design adaptions of website plug-ins. At the same time, we agree with Vermeeren et al. [57] and Law et al. [33] that an exclusive focus on quantitative UX measurement (through, e.g., web analytics metrics) might ignore relevant insights of qualitative measurement approaches. Consequently a balance of various measurement tools and approaches should be promoted. In culturally sensitive development processes, the research and design team can implement UX-focused user tracking to identify suitable levers for country-specific design adaptions. Once significant differences in the usage behaviors for certain steps of a website plug-in have been identified, researchers and developers can, e.g., efficiently set-up subsequent A/B-tests and investigate the impact on the click behavior for different designs. This includes but is not limited to more or less information for such plug-insteps, different designs (colours, fonts, etc.) or simply a different user flow through the plug-in. Changes in the design can then be analyzed through further user tracking and supplemental qualitative evaluations.

In summary, our work was guided by the motivation to pursue a quantitative approach based on web analytics metrics to identify UX-related, country-specific usage behaviors of a website plug-in. We aim to foster an ongoing discussion about cross-cultural UX design as well as a suitable balance between qualitative and quantitative UX measurement - following up on the investigation of large-scale behavioral data. In particular, however, we want to emphasize that the challenging need of globally acting companies to analyze country-specific preferences and usage patterns requires cost-efficient and quickly adaptable UX measurement tools. In this light, we perceive our work as a constructive starting point for further cross-cultural investigations based on large-scale behavioral data.

References

1. Ahtinen, A., Ramiah, S., Blom, J., Isomursu, M.: 2008. Design of Mobile Wellness Applications: Identifying Cross-Cultural Factors. In: Proceedings of the 20th Australasian Conference on Computer-Human Interaction (OZCHI), pp. 164–171. ACM Press (2008)
2. Al-Shamaileh, O., Sutcliffe, A.: Investigating a Multi-faceted View of User Experience. In: Proceedings of the 24th Australian Computer-Human Interaction Conference (OZCHI), pp. 9–18. ACM Press (2012)
3. Andreasen, M., Nielsen, H., Schrøder, S., Stage, J.: What Happened to Remote Usability Testing? An Empirical Study of Three Methods. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1405–1414, ACM Press (2007)
4. Aykin, N: Overview: Where to Start and What to Consider. In: Aykin, N. (Ed.): Usability and Internationalization of Information Technology, pp. 3–20. Lawrence Erlbaum Associates, Inc., New Jersey (2005)
5. Barber, W., Badre, A.: Culturability: The Merging of Culture and Usability (1998) `http://research.microsoft.com/en-us/um/people/marycz/hfweb98/barber/`
6. Bargas-Avila, J., Hornbæk, K.: Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Empirical Studies of User Experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2689–2698. ACM Press (2011)
7. Battarbee, K., Koskinen, I.: Co-experience: user experience as interaction. CoDesign 1 (1), 5–18 (2005)
8. Beaton, J., Kumar, R.: Indian Cultural Effects on User Research Methodologies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA), pp. 4267–4271. ACM Press (2010)
9. Bjørneseth, F., Dunlop, M., Strand, J.: Dynamic positioning systems: usability and interaction styles. In: Proceedings of the 5th Nordic Conference on Human-Computer Interaction (NordiCHI), pp. 43–52. ACM Press (2008)
10. Chang, T-H., Yeh, T., Miller, R.: GUI Testing Using Computer Vision. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1535–1544. ACM Press (2010)
11. Cohen, J.: Statistical power analysis for the behavioral sciences. Academic Press, New York (1998)
12. Dunlop, M., Hamilton, I,. Komninos, A., Nicol, E.: Shake 'N' Tap: A gesture enhanced keyboard for older adults. In: Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (MobileHCI), pp. 525–530. ACM Press (2014)
13. Dunlop, M., Mcgregor, B., Elliot, M.: Using smartphones in cities to crowdsource dangerous road sections and give effective in-car warnings. In: Proceedings of the SEACHI 2016 on Smart Cities for Better Living with HCI and UX (SEACHI), pp. 14–18. ACM Press (2016)

14. Frandsen-Thorlacius, O., Hornbæk, K., Hertzum, M., Clemmensen, T.: Non-Universal Usability? A Survey of How Usability is Understood by Chinese and Danish Users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 41–50. ACM Press (2009)
15. Gerea, C., Herskovic, V.: Measuring User Experience in Latin America: An Exploratory Survey. In: Proceedings of the Latin American Conference on Human Computer Interaction (CLIHC). ACM Press (2015)
16. Gordon, M., Ouyang, T., Zhai, S.: WatchWriter. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 3817–3821. ACM Press (2016)
17. Gorman, T., Rose, E., Yaaqoubi, J., Bayor, A., Kolko, B.: Adapting Usability Testing for Oral, Rural Users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1437–1440. ACM Press (2011)
18. Grubbs, F.: Procedures for detecting outlying observations in samples. Technometrics 11 (1), 1–21 (1974)
19. Hassenzahl, M., Tractinsky, N.: User experience - a research agenda. Behaviour & Information Technology 25 (2), 91–97 (2006)
20. Hassenzahl, M.: User experience (UX): towards an experiential perspective on product quality. In: Proceedings of the 20th Conference on l'Interaction Homme-Machine (IHM), pp. 11–15. ACM Press (2008)
21. He, Y., Zhao, C., Hinds, P.: Understanding Information Sharing from a Cross-cultural Perspective. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA), pp. 3823–3828. ACM Press (2010)
22. Hofstede, G., Hofstede, G. J., Minkov, M.: Cultures and Organizations. Software of the Mind. McGraw-Hill, New York (2010)
23. Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing. In: IEEE Transactions on Multimedia 16 (2), 541–588 (2014)
24. ISO DIS. 2009. 9241-210. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. International Standardization Organization (ISO). Switzerland (2009).
25. Jewell, C., Salvetti, F.: Towards a Combined Method of Web Usability Testing: An Assessment of the Complementary Advantages of Lab Testing, Pre-Session Assignments, and Online Usability Services. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA), pp. 1865–1870. ACM Press (2012)
26. Jordan, P.: Human factors for pleasure in product use. Applied Ergonomics 29 (1), 25–3 (1998)
27. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.: User Experience Over Time: An Initial Framework. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 729–738. ACM Press (2009)
28. Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., Bonchi, F.: Cultures in community question answering. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 175–184. ACM Press (2015)
29. Komninos, A., Nicol, E., Dunlop, M.: Designed with Older Adults to SupportBetter Error Correction in SmartPhone Text Entry. In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI), pp. 797–802. ACM Press (2015)
30. Lachner, F., Nägelein, P., Kowalski, R., Spann, M., Butz, A.: Quantified UX: Towards a Common Organizational Understanding of User Experience. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI). ACM Press (2016)

31. Lallemand, C., Gronier, G., Koenig, V.: User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. Computers in Human Behavior 43, 35–48 (2015)
32. Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., Kort, J.: Understanding, scoping and defining user experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 719–728. ACM Press (2009)
33. Law, E., Van Schaik, P., Roto, V.: Attitudes towards user experience (UX) measurement. International Journal of Human Computer Studies 72 (6), 526–541 (2014)
34. Lindley, S., Wallace, J.: Placing in Age: Transitioning to a New Home in Later Life. Transactions on Computer-Human Interaction 22 (4), 1–40 (2015)
35. Lindley, S.: Making Time. In: 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW), pp. 1442–1452. ACM Press (2015)
36. Malinen, S., Nurkka, P.: The role of community in exercise: Cross-cultural study of online exercise diary users. In: Proceedings of the 6th International Conference on Communities and Technologies (C&T), pp. 55–63. ACM Press (2013)
37. Marcus, A., Gould, E.: Crosscurrents: Cultural Dimensions and Global Web User-Interface Design. Interactions 7 (4), 32–46 (2000)
38. Mekler, E., Hornbæk, K.: Momentary Pleasure or Lasting Meaning ? Distinguishing Eudaimonic and Hedonic User Experiences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 4509–4520. ACM Press (2016)
39. Nielsen, J.: Designing for International Use (Panel). In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 291–294. ACM Press (1990)
40. Norman, D.: Emotional design. Ubiquity 45 (4), 1–1 (2004)
41. Obrist, M., Roto, V., Väänänen-Vainio-Mattila, K.: User Experience Evaluation - Do You Know Which Method to Use? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2763–2766. ACM Press (2009)
42. Obrist, M., Wurhofer, D., Gärtner, M., Förster, F., Tscheligi, M.: Exploring children's 3DTV experience. In: Proceedings of the 10th European conference on Interactive tv and video (EuroiTV), pp. 125–134. ACM Press (2012)
43. Oliveira, N.: Culture-aware Q&A Environments. In: Proc. DC CSCW Companion, pp. 101–104. ACM Press (2015)
44. Oliveira, N., Andrade, N., Reinecke, K.: Participation Differences in Q&A Sites Across Countries: Opportunities for Cultural Adaptation. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI). ACM Press (2016)
45. Rea, L. M., Parker, R. A.: Designing and conducting survey research: A comprehensive guide. John Wiley & Sons (2014).
46. Reinecke, K., Bernstein, A.: Improving Performance, Perceived Usability, and Aesthetics with Culturally Adaptive User Interfaces. Transactions on Computer-Human Interaction 18 (2), 1–29 (2011)
47. Reinecke, K., Bernstein, A.: Predicting User Interface Preferences of Culturally Ambiguous Users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA), pp. 3261–3266. ACM Press (2008)
48. Reinecke, K., Gajos, K.: Quantifying Visual Preferences Around theWorld. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 11–20. ACM Press (2014)
49. Reyal, S., Zhai, S., Kristensson, P.: Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 679–688. ACM Press (2015)
50. Rodden, K., Hutchinson, H., Fu, X.: Measuring the User Experience on a Large Scale: User-Centered Metrics for Web Applications. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2395–2398. ACM Press (2010)

51. Rödel, C., Stadler, S., Meschtscherjakov, A., Tscheligi, M: Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. In: Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI), pp. 1–8. ACM Press (2014)
52. Roto, V., Law, E., Vermeeren, A., Hoonhout, J.: UX White Paper. Bringing clarity to the concept of user experience. Result from Dagstuhl Seminar on Demarcating User Experience (2011)
53. Sun, H.: Building a Culturally-Competent Corporate Web Site: An Exploratory Study of Cultural Markers in Multilingual Web Design. In: Proceedings of the 19th annual international conference on Computer documentation (SIGDOC), pp. 95–102. ACM Press (2001)
54. Swallow, D., Blythe, M., Wright, P.: Grounding Experience: Relating Theory and Method to Evaluate the User Experience of Smartphones. In: Proceedings of the 2005 annual conference on European association of cognitive ergonomics (EACE), pp. 91–98. ACM Press (2005)
55. Tuch, A., Trusell, R., Hornbæk, K.: Analyzing Users' Narratives to Understand Experience with Interactive Products. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2079–2088. ACM Press (2013)
56. Tullis T., Albert, B.: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Second Edition. Morgan Kaufmann, Waltham (2013)
57. Vermeeren, A., Law, E., Roto, V., Obrist, M., Hoonhout, J., Väänänen-Vainio-Mattila, K.: User experience evaluation methods: current state and development needs. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction (NordiCHI), pp. 521–530. ACM Press (2010)
58. Walsh, T., Nurkka, P., Walsh, R.: Cultural Differences in Smartphone User Experience Evaluation. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia (MUM), pp. 1–9. ACM Press (2010)
59. Walsh, T., Nurkka, P.: Approaches to Cross-Cultural Design: Two Case Studies with UX Web-Surveys. In: Proceedings of the 24th Australian Computer-Human Interaction Conference (OZCHI), pp. 633–642. ACM Press (2012)
60. Walsh, T., Vainio, T.: Cross-Cultural Design for mHealth Applications. In: Proceedings of the 23rd Australian Computer-Human Interaction Conference (OZCHI). ACM Press (2011)
61. Yatani, K., Novati, M., Trusty, A., Truong, K.: Review Spotlight: A User Interface for Summarizing User-generated Reviews Using Adjective-Noun Word Pairs. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1541–1550. ACM Press (2011)