Less Annoying: Quality of Experience of Commonly Used Mobile Applications

Alexandre De Masi
University of Geneva
Geneva, Switzerland
alexandre.demasi@unige.com

Katarzyna Wac
University of Geneva
Geneva, Switzerland
katarzyna.wac@unige.ch

ABSTRACT

In recent years, research on the Quality of Experience (QoE) of smartphone applications has received attention from both industry and academia due to the complexity of quantifying and managing it. This paper proposes a smartphone-embedded system able to quantify and notify smartphone users of the expected QoE level (high or low) during their interaction with their devices. We conducted two in the wild studies for four weeks each with Android smartphones users. The first study enabled the collection of the QoE levels of popular smartphone applications' usage rated by 38 users. We aimed to derive an understanding of users' QoE level. From this dataset, we also built our own model that predicts the QoE level for application category. Existing QoE models lack contextual features, such as duration of the user interaction with an application and the user's current physical activity. Subsequently, we implemented our model in an Android application (called expectQoE) for a second study involving 30 users to maximize high QoE level, and we replicated a previous study (2012) on the factors influencing the QoE of commonly used applications. The expectQoE, through emoji-based notifications, presents the expected application category QoE level. This information enable the user's to make a conscious choice about the application to launch. We then investigated whether if expectQoE improved the user's perceived QoE level and affected their application usage. The results showed no conclusive user-reported improvement of their perceived QoE due to expectQoE. Although the participants always had high QoE application usage expectations, the variation in their expectations was minimal and not significant. However, based on a time series analysis of the quantitative data, we observed that expectQoE decreased the application usage duration. Finally, the factors influencing the QoE on smartphone applications were similar to the 2012 findings. However, we observed the emergence of digital wellbeing features as facets of the users' lifestyle choices.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Quality of Experience, Smartphone, Notification, In The Wild, Expectations, Mobile Applications

ACM Reference Format:
Alexandre De Masi and Katarzyna Wac. 2022. Less Annoying: Quality of Experience of Commonly Used Mobile Applications. In *13th ACM Multimedia Systems Conference (MMSys '22), June 14–17, 2022, Athlone, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3524273.3528183

1 INTRODUCTION

Smartphones are an integral part of modern life. They enable users to access online services to communicate and exchange information around the world. They allow them to create or consume content in different contexts; however, the experience can be impacted by the smartphone user's context. This context often changes due to circumstances such as the physical activity of the user, varying the smartphone application user experience as a result [13]. As such, the term Quality of Experience (QoE) was coined to mirror the known Quality of Service (QoS, [37]) concept from the telecommunication and networking domains. Whereas the QoS only focuses on quantitative information. The QoE measurement is an expansion of the QoS and includes qualitative information related to the experience itself, and thus prioritizes the end user. QoS focused on metrics obtained on user-end device and networking device (e.g., jitter, amount of packet error and dropped) which transport content. Contrary to QoE, which focuses on the experience encompassed in the content. The QoE is defined by the Qualinet White Paper [23] as "an application or service user's degree of delight or annoyance".

Many previous works [7, 10, 16] have only focused on quantifying the smartphone applications and web browsing QoE based on QoS metrics within laboratory settings where the authors simulated external factors (e.g., a bandwidth limitation or reduced video bitrate), missing important contextual factors such as user's habits and current activity. To bridge this gap, we first aimed to quantify the QoE of smartphone applications. As such, we employed a mixed-methods approach and collected application usage QoE ratings via an in the wild study (S1) with 38 participants over four weeks. During this study, we deployed an Android logger application, named mQoL-Lab [4], that collected context information. The users had the opportunity to rate their application usage through Ecological Momentary Assessment (EMA; [34]). From the collected dataset, we built a QoE classification model that predicts the application category QoE level between two labels: high or low. The labels correspond to the level of acceptability from the end-user perspective [32]. The QoE classification model is based on features from three different perspectives: user (e.g., intent to accomplish), system (e.g.,

MMSys '22, June 14–17, 2022, Athlone, Ireland
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9283-9/22/06…$15.00
https://doi.org/10.1145/3524273.3528183

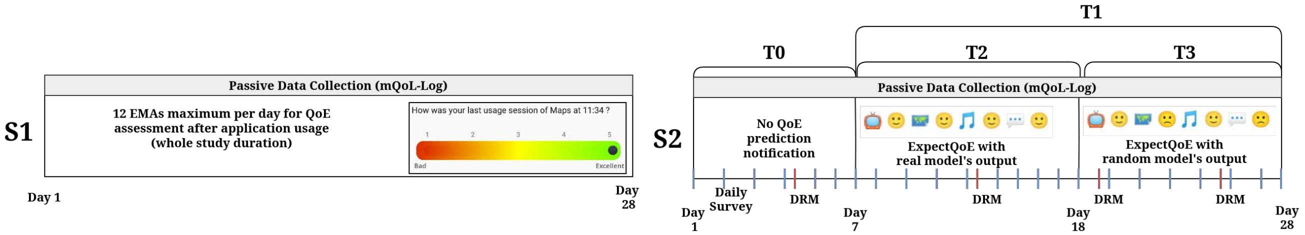
S1
Passive Data Collection (mQoL-Log)
12 EMAs maximum per day for QoE assessment after application usage (whole study duration)
How was your last usage session of Maps at 11:34 ?
1
2
3
4
5
Bad
Excellent
Day 1
Day 28
S2
T1
T0
T2
T3
Passive Data Collection (mQoL-Log)
No QoE prediction notification
ExpectQoE with real model's output
ExpectQoE with random model's output
Daily Survey
DRM
DRM
DRM
DRM
Day 1
Day 7
Day 18
Day 28

Figure 1: Study S1 and S2 Timeline and Research Methods

QoS metrics) and context (e.g., user physical activity). Related previous works only went as far as estimating the QoE level of video or social media applications on smartphones [8], to help telecommunication providers offer better network conditions through core network parameters. This information was often processed after the study and thus retained from the end users.

Moreover, the telecommunication providers continuous upgrade of network equipment leading to a better user experience will always be limited by the users' current smartphone hardware, the network protocols, and the physics of wireless broadband. There are techniques to reduce user annoyance based on hardware technology upgrades or architectural system design (e.g., microservices and edge computing). However, these techniques could fail when the service provider and the network are inaccessible, or when the user's intent is unpredictable and requires real-time access to the content (i.e., the content is impossible to cache). Hence, software approaches based on human-computer interaction (HCI) technics could be a potential solution which have not been employed to grand extend yet. One of the solution in place is the indicator on smartphone which always presents the user's network state. Accordingly, users expect internet-enabled applications to be slow when no bars are shown.

We conducted our study with the hypothesis that an intervention approach could influence smartphone users' behaviour. The users would attempt to avoid annoying experiences and limit their application usage duration. Previous studies have shown that notifications are capable of influencing smartphone users by communicating information about the intervention topic [25, 26, 30] (e.g. reducing exposure to low interest notification or pushing user to engage less in certain games). Thus, we approached maximizing users' QoE by providing notifications that aimed at limiting their exposure to applications with a low predicted QoE level. We implemented our QoE level prediction model (expectQoE) into our mQoL-Lab and conducted a second four-week study (S2) with 30 participants to investigate the model's influence with notifications to present to the user the predicted QoE level.

Besides QoE models, we focus on understanding the factors influencing users' experiences as they vary through context and time. The 2012 study of Ickin et al. [18] on this subject identified these factors through a user study. Hence, we replicated part of their work for S2. We focused on previously defined factors influencing the QoE of smartphones and the factors current evolution.

2 RELATED WORK

Assessing users' perceived experience of smartphone applications in the wild has been performed by Casas et al. [7, 9]. The researchers focused on the QoE of smartphone applications in cellular networks. They labeled their data in the field, but their participants were instructed to accomplish a specific task. Such study design can impact participants' annotation process. In a later work, the authors modeled smartphone application QoE [6] with success (95% accurate), although they did not deploy or test their models outside a laboratory setting. Furthermore, they limited their focus to video and audio streaming only.

Schwind et al. [33] used the MONROE [1] hardware platform to collect network and audio and video streaming metadata from online services on public transport (trains and buses) nodes in European countries. The study focused on video streaming and modeling the QoE based on video streaming bitrates. However, this work did not represent real user and smartphone interactions but only the results of multiple network tests in different contexts (i.e., mobility induced cell tower changes). Other researchers attempted to measure and predict network quality on trains [22] by measuring QoS metrics. However, they did not factor the user's context, the application used, or previous user's experience into their models. Moreover, their model was based on a dataset collected on a specific train ride. Summarizing, the previous listed works focused on QoE estimation only.

We propose to use intervention to limit the smartphone users' exposure to annoying application experience. Interventions on smartphones are primarily present in health-based research. For instance, a smartphone-based intervention showed success in motivating physical activity in a student population [29], to change the participants' behavior through notifications [28]. Notifications are considered as an intervention tool due to their unexpected nature and informational content [26]. Additionally, smartphone-based interventions studies have employed notification to promote digital wellbeing [27, 40] with some success. While QoE modeling has been researched for smartphone applications, maximizing the QoE through notifications has not. Our work addresses this research gap, offering a solution via a notification to reduce user annoyance preemptively.

3 METHODOLOGY

3.1 Study Protocols (S1, S2)

Figure 1 depicts our approach and the study protocol for S1 (2018) and S2 (2021). As defined, S1 was for building the QoE model, while

S2 was for model validation. The participant's information is presented respectively in Section 3.2.1 for S1 and Section 4.1 for S2. The participants were recruited on the university's campus via flyers, mailing lists, and social network posts. We focused on Android smartphone users who had lived at least five years in the Great Geneva region, at the border of France and Switzerland. Participants were required to use a minimum of four target applications daily, and a maximum of all, and have the most recent OS version on their smartphones. We focused on popular smartphone applications: Instagram, WhatsApp, Spotify, Facebook, Chrome, Facebook Messenger, and Google Maps. These applications were selected due to their high installation number from the Google Play Store and their previous selection in past work [7]. Once selected, the participants were invited to our laboratory for a demonstration of the logger and of the tasks to accomplish during the studies. Each participant gave consent before enrolling in the study and downloading the mQoL-Lab application. This Android logger was employed in both studies to collect passive information on the user context. Tables 1 presents the data collected during S1 and the data used for estimating the QoE level during S2. Since S2 was more recent (2021 versus 2018), the availability of the data had changed. Previously available information like low-level network statistics were removed. mQoL-Lab also enables the collection of smartphone usage QoE level ratings (i.e., via EMAs) in situ [12]. In S1, the participants annotated the QoE level of their last application usage through a 5-point Mean Opinion Score (MOS) [19]. This method was previously used in QoE smartphone studies [9, 18, 31]. An EMA was triggered by an application events (i.e., closing or switching). The question "How was your last usage session of {'app name'} at {'time'}?" enabled the participants to know what application they were rating. The MOS scale contained the followings scores: poor (1), bad (2), fair (3), good (4), and excellent (5). The EMA also contained multiple choice questions about the intent the user wished to accomplish with the application: consume content, share or create content, read text message, write text message, control an app (e.g., start or stop music), video call, or audio call. We limited the number of EMAs to 12 per day during waking hours between 7:00 and 21:00, and included a 20-minute timeout between consecutive EMAs to reduce the study burden. We categorized the obtained data in the studies into three perspectives (Table 1) (i) User centric: qualitative data obtained from the user; (ii) System centric: quantitative data obtained from smartphone sensors linked to the smartphone hardware and software state (e.g., network QoS); (iii) Context centric: quantitative data obtained from smartphone sensors and characterized by a strong in-situ nature.

The protocol for S2 differed, as shown in Figure 1. S2 timeline is composed of three distinct periods, T0 was the baseline period, it contains the participants' application usage habits (i.e., passive collection). Then T1 marked the beginning of the intervention period with expectQoE. T1 was composed of two periods, T2 in which the participants were notified with the real output of our QoE model. Contrary to T3, in which the participants received random QoE level. The periods in the S2 timeline enable us to capture the supposed influence of expectQoE on the participant's application usage. T0 represents the baseline data. T1 focus on the influence of expectQoE across T2 (real model output) and T3 (random output).The predicted QoE levels in T2 were estimated using our model built from the data collected in S1. S2 study also included three types of questionnaires [18]: weekly Day Reconstruction Method (DRM) [21], semi-structured interviews, daily online QoE surveys at the end of the day at a random time between 19:30 and 20:00, and notifications that presented the expectQoE level of the four application categories which the participants could rate (12 maximum per day from 8:00 to 21:00, 5 minutes minimum between two notifications). The question was asked as follows: "Did your application usage sessions meet your expectations?". A slider was used to answer from 1 (not at all) to 5 (enormously). The weekly remote DRM interviews were conducted to discuss their previous 24 hours of smartphone usage and any other recent annoying experience on participants' smartphones. With the daily surveys, we queried the participants about their overall QoE, expectations, stress level, and usefulness of the expectQoE system. We used an MOS scale from 1 to 5 for each indicator except for the stress level, which used a scale from 0 to 10 [3]. For S2 study, we developed a notification-based EMA. The expectQoE presented the application categories and predicted QoE level via emojis. Emojis surrogate plain text [24] and have been heavily used to generate notification-to-application interactions [36]. We mapped each application category to a specific emoji: communication and social 💬, music and audio 🎵, video 📺, and travel and local 🗺️. In notification buttons were used to provide feedback with the thumbs-up 👍 and thumbs-down 👎 emoji. The QoE level was indicated via the slightly-smiling-face emoji 🙂 for a predicted high-level QoE and the worried emoji 😟 for a predicted low-level QoE. The expectQoE notification content contained the dyad "category, predicted QoE" concatenated for all the categories in an emoji sequence (e.g., Figure 1). To preemptively limit participants' fatigue from seeing the same content in the notification area, we randomized the order of the four dyads. The notification was triggered after an application usage started and disappeared once the user rated it. The emojis were not updated due to a limitation from the Android notification's nature. Each update would have created a new notification which could have visually disturbed the participant. Moreover, they were only available beginning Day 7 (T2) of the study. Furthermore, after Day 18 (T3) of the study, the

Table 1: Information Collected per Study (mQoL-Lab)

Perspective	Domain	Raw Features Available	S1	S2
User	Intent	Intent the user tried to accomplish by launching an application	✓	
	Application	Name of the application launched by the user	✓	✓
	Session	Duration of the application usage session (ms)	✓	
System	Network	Cell and Wi-Fi signal strength, Wi-Fi speed, Cellular up and down bandwidth,	✓	✓
		Active ping test to measure the round-trip time to the University server, handover, IP version, aggregated traffic packet statistics	✓	
		Netstats (i.e., TCP states per socket)	✓	
	Battery	Energy level from the battery (capacity in %)	✓	✓
Context	Physical Activity	User's physical activity from Google Activity Recognition (walking, running, still, on bicycle, in vehicle)	✓	✓

QoE levels presented were randomized, thus enabling us to test the impact of our model (T2) versus a balanced distribution of random QoE level as output (T3, assumed 50%/50% for low/high QoE).

3.2 Modeling QoE (S1)

3.2.1 S1 Dataset. In S1, the age distribution of the 38 participants is as follows. Thirteen were young adults (two between 18-20 y.o. and eleven between 21-29 y.o.), followed by ten participants between 30-39 y.o., two between 40 and 49 y.o., two participants between 50-59 y.o. and two non-disclosed. The gender distribution is as follows: fifteen were women, twenty-one men and one non-disclosed. Furthermore, the participants' education level (i.e., last successful diploma obtained) was as follow, and five participants had a PhD degree, followed by fifteen participants who had a master's degree, then four had a bachelor's degree, twelve had a high school diploma or equivalent, and finally two participants had no diploma.

We collected 6,308 ratings (166 ± 89 per participant) of application usage QoE. Only five participants triggered the maximum possible number of EMAs. At the end of the study, the participants answered 75±2% of the triggered EMAs. In general, the participants rated their QoE as good (MOS > 4). The QoE ratings were mapped into two groups: high and low in accordance with Schatz et al.'s user accessibility threshold [32], and due to the ratings' imbalance. More than two categories would lead to an even more imbalanced dataset than binary setting, impacting the model performances. Also, the choice of the threshold's values between the categories would have to be validated. The binary classification approach enables the construction of a robust model metrics wise (i.e., AUC). Ratings higher than or equal to 3.5 were classified as high; all other ratings were classified as low. The prevalence of high QoE levels was 93.5% versus 6.5% for low QoE levels.

3.2.2 ExpectQoE Features. We identified features to build a QoE-level prediction model based on the data collected during S1. Contrary to previous works, we required that the model had to make predictions directly on the devices. Accordingly, some aggregated information was inaccessible due to time constraints (e.g. time-based aggregated feature: application usage duration). We aimed at classifying the QoE levels of the following application categories: communication (e.g., WhatsApp and Facebook Messenger) and social (e.g., Instagram, Twitter, and Facebook), music and audio (e.g., Spotify), video (e.g., YouTube and Netflix), and travel and local (e.g., Google Maps). These categories represent more than 60% of the applications launched on smartphones [5]. We selected other features that were accessible on-the-fly on Android 12. The selected features were presented in Table 1 column labeled S2.

3.2.3 ExpectQoE Building. The model we applied follows the on-device model construction presented by De Masi et al. [14]. Our model could be subject to overfitting since the S1 dataset contains a higher amount of high-level QoE annotation than low. Hence, we undersampled the S1 dataset, resulting in maintaining 386 samples for each class. We split the dataset into a training set (70%) and a testing set (30%). We applied ten-fold cross-validation and trained the model with the XGBoost algorithm [11], which has been proven to perform efficiently with tabular data. We repeated the same process 10 times. During each fold, the undersampling selected different samples from the majority class. We obtained an average Area Under the Curve (AUC) 75 ± 7% (higher is better) on the test dataset to classify the QoE level (high/low). We exported the model with the highest AUC (82%) into our logger application for S2.

4 EXPECTQOE EVALUATION AND USER STUDY RESULTS (S2)

4.1 Demographic: S2

The age distribution of the 30 participants is as follows. Four were young adults (two between 18-20 y.o. and 11 between 21-29 y.o.), followed by eleven participants between 30-39 y.o., one between 40 and 49 y.o., one participant between 50-59 y.o. and one non-disclosed. The gender distribution is as follows. Nine were women, followed by nineteen men and two non-disclosed. Furthermore, the participants' education level is as follows: four participants had a PhD degree, followed by nine participants who had a master's degree, then seven had a bachelor's degree. Finally, seven participants had a high school diploma or equivalent, and only one participant did not have any degree. Two participants chose to not answers this question.

4.2 Factors Influencing QoE

In 2012, Ickin et al. [18] ran a study with 29 Android smartphone users for four weeks focusing on understanding smartphone QoE. The authors employed the DRM method to analyze the relations and causality between QoE annotations collected during the study, QoS, and context. Two independent coders clustered the terms with the most affinity. In the end, they distinguished seven factors influencing QoE. We used the seven factors as a template during our S2 weekly interviews. Overall, we collected 120 expressions from the 30 participants. Two researchers familiar with the QoE and smartphones domain coded the expressions, and the measure of agreement was greater than 96%. Overall, we found similarities with the past work, yet the factor meanings have changed with time.

4.2.1 Application interface design. The application's interface design was commented on often. The participants enjoyed the interface of the notification bar. Contrary to [18], the participants complained about the content of the applications versus their mood at that time (e.g., announcement of the death of a family member via an application).

4.2.2 Application performance. Twenty-four participants reported problems with sharing photos and streaming videos for example: "the videos were not loading" (P22), "it's problematic, the connection is bad, on YouTube I have to wait a lot for a video to load" (P26). The participants were also conscious of the capacity of the network to which they were connected. In particular, twenty-three participants commented on roaming between countries and the time needed for their smartphone to connect to a new network. One participant experienced low QoE due to network roaming problems (P15). Two participants had to set up the cell network manually due to their proximity to a foreign cell tower (at the border). Only three participants reported playing video games on their devices. Overall, the participants were able to discern whether the performance of an application was due to the application itself or to an underlying

Table 2: S2: Participants Ratings to ExpectQoE Predictions

Participant	QoE Level	Response	Answered	Triggered	👍/👎	👍/👎	👍/👎
	T0 [%]	T1	T1	T1	T1 [%]	T2 [%]	T3 [%]
ID	High/Low	Rate [%]	[n]	[n]	S2 Total [%]	Model [%]	Random [%]
0	0.66/0.34	86	19	22	0.26/0.74	0.25/0.75	0.27/0.73
1	0.75/0.25	80	202	252	0.49/0.51	0.76/0.24	0.19/0.81
2	0.77/0.23	48	121	252	0.66/0.34	0.62/0.38	0.89/0.11
3	0.96/0.04	48	122	252	0.48/0.52	0.38/0.62	0.63/0.37
4	0.82/0.18	69	24	35	1.00/0.00	1.00/0.00	0.00/0.00
5	0.98/0.02	79	200	252	0.99/0.01	0.99/0.01	0.00/0.00
6	0.89/0.11	73	185	252	0.53/0.47	0.54/0.46	0.52/0.48
7	0.86/0.14	69	173	252	0.32/0.68	0.74/0.26	0.10/0.90
8	0.89/0.11	87	219	252	0.57/0.43	0.57/0.43	0.57/0.43
9	0.69/0.31	90	226	252	0.49/0.51	0.35/0.65	0.66/0.34
10	0.92/0.08	98	118	120	0.98/0.02	1.00/0.00	0.98/0.02
11	0.92/0.08	87	219	252	0.97/0.03	0.96/0.04	0.97/0.03
12	0.99/0.01	62	155	252	0.56/0.44	0.70/0.30	0.44/0.56
13	0.95/0.05	77	59	77	0.49/0.51	0.56/0.44	0.40/0.60
14	0.96/0.04	99	128	129	0.23/0.77	0.35/0.65	0.19/0.81
15	0.90/0.10	60	131	220	0.43/0.57	0.79/0.21	0.06/0.94
16	0.91/0.09	100	252	252	0.98/0.02	0.98/0.02	0.99/0.01
17	0.88/0.12	41	104	252	0.99/0.01	1.00/0.00	0.98/0.02
18	0.89/0.11	54	135	252	0.72/0.28	0.72/0.28	0.00/0.00
19	1.00/0.00	92	88	96	0.9/0.1	0.89/0.11	0.91/0.09
20	0.9/0.1	78	197	252	0.59/0.41	0.57/0.43	0.64/0.36
21	0.92/0.08	98	135	138	0.16/0.84	0.19/0.81	0.11/0.89
22	0.9/0.1	52	92	178	0.36/0.64	0.48/0.52	0.12/0.88
23	0.83/0.17	100	252	252	0.56/0.44	0.54/0.46	0.60/0.40
24	0.97/0.03	86	96	111	0.45/0.55	0.60/0.40	0.34/0.66
25	0.82/0.18	60	152	252	0.47/0.53	0.77/0.23	0.32/0.68
26	1.0/0.0	97	68	70	0.96/0.04	0.92/0.08	1.00/0.00
27	0.92/0.08	98	52	53	0.81/0.19	0.84/0.16	0.75/0.25
28	0.9/0.10	65	165	252	0.83/0.17	0.88/0.12	0.78/0.22
29	0.92/0.08	86.0	217.0	252.0	0.18/0.82	0.14/0.86	0.23/0.77
ALL	0.88/0.11	77±18	144±65	193±82	0.61/0.29	0.67/0.33	0.48/0.52

network problem. That is different from the 2012 study, where such distinction was not made by the participants.

4.2.3 *Battery.* The batteries were able to sustain the smartphones for more than a day with high utilization from the participants. However, one participant reported carrying an extra battery when travelling in another country as their smartphone was used to guide their group and thus consumed more energy due to using GPS (P26).

4.2.4 *Phone features.* The participants reported enjoying the camera quality. Four participants used the hotspot function to share through their 4G connection with their friends or other devices (e.g., laptop or game console) when their home Internet delivered low QoE.

4.2.5 *Applications and data connectivity cost.* More than half the participants (17) mentioned having an unlimited mobile data subscription, hence their use of the hotspot feature. Overall, they were satisfied with the free applications available. However, four participants paid for application subscriptions that enhanced the application features. Contrary to [18], when this services were not available on smartphone, seven participants subscribed to multiple streaming services. Overall, we found that Spotify (11 participants) and Netflix (11 participants) were among the most used services.

4.2.6 *Routine.* Twenty-five participants reported following an identical routine in the morning and in the evening. In both cases, they used a set of applications, often communication (e.g., WhatsApp, email), before starting or finishing their day, which corresponds to the findings of [18] regarding user routines.

4.2.7 *Lifestyle.* We observed a trend in the lifestyle factor that was not seen in 2012 [18]. Namely, participants limited their interaction with smartphones during work and at night. We identified four levels of this digital wellbeing behavior: (i) Smartphone physically inaccessible (1 participant): The device is placed outside the bedroom at night, limiting accessibility; (ii) Plane mode (3 participants): All access to any network is disabled; (iii) Data network off (4 participants): The Wi-Fi and cell data access are turned off; the smartphone user can still receive a call, but Internet applications are unavailable; (iv) Smartphone enabled (7 participants): The "Do not disturb" mode stops all notifications from appearing on the screen, removes its signaling modalities (vibration and sound), and limits application usage based on time: "I put the phone on silence when I arrive at work" (P20). The applications run and are synchronized in the background. Additionally, contrary to past work, we found that many participants used more diverse applications that supported their lifestyle such as finance, spirituality, and health applications.

4.3 ExpectQoE Model Evaluation Dataset

Overall, we collected 64,179 application usage sessions from the thirty participants with 464 unique applications. The participants used mostly communication applications throughout the study which corresponds to the findings of Bohmer et al. [5] at the beginning of the smartphone revolution. The ten most launched applications by the participants were WhatsApp (13.7%), Instagram (7.4%), Chrome (6.1%), Telegram (6%), Snapchat (3.8%), Gmail (3.7%), Phone dialer (2.9%), YouTube (2.9%), Message (SMS and MMS, 2.4%) and Facebook (2.3%). These applications represent 52% of the total launch application during S2. The top ten applications in which participants spend the most time correspond partially to the ten most launched applications [5].

The Table 2 summarized the participants interactions with expectQoE in T1. Overall, they triggered on average 193±82 expectQoE notifications due to their application usage. Only twenty participants used their smartphones enough to trigger the maximum amount of expectQoE notifications (252) throughout the study, impacting the amount of rating collected. However, the participants rated expectQoE on average 144±65 times. The categories for which expectQoE provided the QoE level prediction in T2 and T3 were communication and social, music and audio, travel and local and video player. Overall, these categories represent on average 65 ± 15% of the total application usage of S2 participants (T0+T1).

We examined the distribution of the expectQoE predictions for the real model output (T2) and the random model (T3). Overall, we observed that during the random model period (T3), the high and low QoE-level predictions were equally distributed among the participants (high QoE level: 49±5%; low QoE level: 51±5%). However, when the real model predicted the QoE (T2), the standard deviation was much higher (high QoE level: 48±24%; low QoE level: 52±24%), indicating a high variation in QoE level for the participants during T2. As such, eleven participants had a low QoE level distribution higher than 60%, contrary to only eight participants with a high QoE level with the same threshold.

4.4 QoE Model Performance (T2)

To investigate the validity of the QoE model predictions, we compare the daily reported QoE from the participants against the model's aggregated output per day in T2.

We employ the Kolmogorov-Smirnov test [2], a non-parametric and distribution-free test. We found that for twenty participants the predicted QoE distribution and reported daily QoE were similar (p<0.04). Indicating that over the day, the model prediction is consistent with the real the participants' feedback. However, for ten participants theses distributions are not statistically significant (p>0.7); the model output does not match their experience.

4.5 ExpectQoE Ratings Analysis

Table 2 presents the ratings given by the participants for each expectQoE notification they received during S2. Overall, the participants rated the expectQoE prediction an average of 77±18% of the time. Their ratings were more positive (thumbs-up, 61%) than negative (thumbs-down, 29%) during the total duration of S2. Only 13 participants rated the expectQoE notification more negatively. This can be explained by the participants' expectation or experience were different from expectQoE notification (i.e., expectQoE misclassified the QoE level).

We compared the ratings for two distinct periods T2 and T3. We observed a higher mean of thumbs-up (67 ± 25%) during T2 than during T3 (48 ± 35%). Both ratings in T2 and T3 are normally distributed (T2 $p < 0.033$, and T3 $p < 0.006$). However, three participants did not provide any ratings during T3 and their data were discarded. Then, we applied a Student's t-test [15] to verify the statistical significant of this difference, and we found $p < 0.025$; hence, we affirmed that the model performance (T2) was better than a random process (T3) from the end user point of view.

4.6 ExpectQoE Notification Effectiveness

Above we analyzed the answers from our participants regarding the notification's effectiveness. However, a quantitative analysis of application usage collected by the logger could assess it better. Therefore, we employed the Multiple Convergent Cross Mapping (MCCM; [39]) from van Berkel et. al. to analyze the causality between the notifications from expectQoE and smartphone use in terms of application usage duration. This method was developed to better understand the interactions between users and technology over time by differentiating causality from correlation based on quantitative data obtained in the wild. MCCM is built on the Convergent Cross Mapping (CCM; [35]) method and extended from the ecology domain based on empirical dynamic modelling methods (EDM; [41]). EDM allows conceptualizing multiple users' behavior as complex nonlinear dynamical systems. CCM is used to investigate the causal relationship between two variables in a time series (i.e., which variable drives the other one) from a complex system. We looked for a positive convergence of the CCM values to determine whether the values were above direct correlation (difference in correlation greater than 0, means the correlation is significant). Then, we were able to establish which variable had a stronger effect by looking at how well one variable forecasts the other. By doing so, we established the direction of causality between the two variables. MCCM enables the aggregation of multiple CCM analyses graphically, processed with each participant time series, by comparing the difference in correlation and asymptotes (convergence point between the two variables forecast) from the CCM results indicating the effect size.

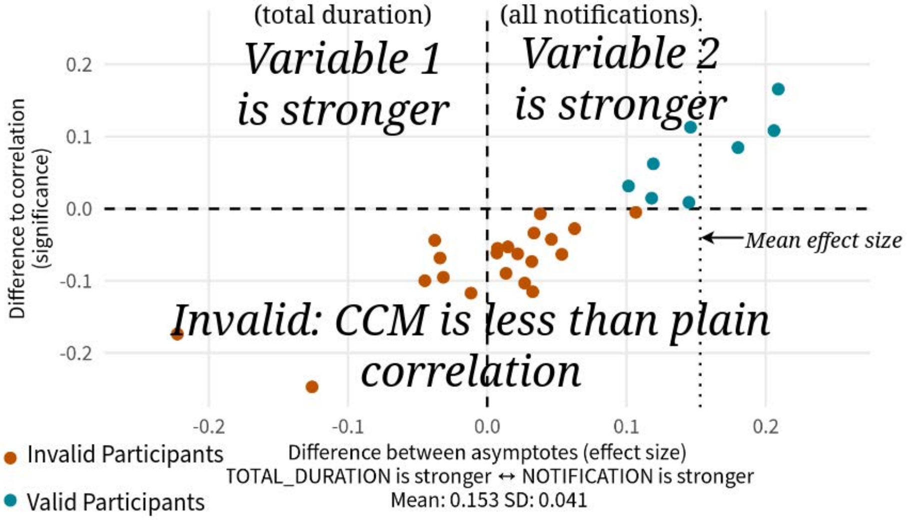

Figure 2: S2: Effect size and Amount of Valid Analysis for Causality Between Application Usage Duration and ExpectQoE Notifications During T1

Table 3: S2: Causality Between Application Usage Duration and expectQoE Notifications

	expectQoE Notification: Effect Size (and STD)		
	All	High	Low
T2 Model	7/30 Valid 0.197 (0.14)	5/30 Valid 0.053 (0.15)	7/30 Valid 0.107 (0.098)
T3 Random	9/30 Valid 0.214 (0.139)	8/30 Valid 0.155 (0.116)	10/30 Valid 0.158 (0.088)

We repeated the MCCM analysis for two time periods (T2, T3) during which we tested three expectQoE notification-based variables (all QoE, high only, low only) against the participants' application usage durations. Figure 2 presents the MCCM visualization. Each point represents a participant, and only the blue points (for valid analysis of a participant) are used to compute the mean effect size. The orange points indicate participants for whom the MCCM analysis failed. The thin dashed line corresponds to the mean effect size. Figure 2 shows that the notification received during the entire S2 duration had an impact on application usage duration for eight participants of thirty. The "invalid" participants' MCCM analysis are explained by the auto-correlated nature of their data (i.e., the application usage duration and the notifications are correlated, but not significant). Thus, the MCCM results are not exploitable on their dataset.

We present the overall results in Table 3, we include the amount of valid participants for each analysis. We observe that expectQoE had an impact on the application usage duration since the effect size is positive (Figure 2 X-axis). The effect size is overall higher with the random model (T3) than with the real model (T2). However, for both models, the expectQoE notifications with low QoE level have a stronger driving effect (0.107 and 0.158) than the high QoE level (0.053 and 0.155). The effect size is stronger in T2 for low QoE level comparing to high QoE level. Hence, the notification impacted more the the application usage duration in T2. Overall, we observed a decrease in application duration usage from 33.7 ± 8[s] in T0 to 28 ± 8[s] in T2.

4.7 Expectation Impact

We explored the expectQoE notification's influence on the participants' daily application usage expectations (in T0, T2 and T3). The average reported answer was 4 ± 0.49 overall, and 4 ± 0.04 in T0, 3.8 ± 0.05 in T2 and 4 ± 0.06 in T3. We focused the analysis on the participant's mean ratings during three different periods: before any use of expectQoE (T0), during expectQoE use with the real QoE model outputs (T2) and finally during expectQoE use with the random model outputs (T3). A one-way Analysis of Variance (ANOVA) of the reported satisfaction expectation ratings was carried out for each period to test if expectQoE influenced the participants' satisfaction. The distribution of the ratings within each period is normally distributed (T0: $p < 0.006$, T2: $p < 0.05$, T3: $p < 0.005$). The results show that the period (T0, T2, T3) has an impact on the participant's ratings, with $F(3, 90) = 3.57$, $p < 0.03$. A Tukey post-hoc test (by setting the $\alpha = 0.05$) revealed that the participant's satisfaction increased significantly from T0 to T2. However, there is no statistically significant difference between the other two pairs of periods (T0, T3 and T2, T3). Hence, we conclude that expectQoE output during T2 has a significant impact on the participant expectations.

4.8 Impact of ExpectQoE on Application Usage Duration

The MCCM results have shown a partial driving force from the expectQoE system on the application usage duration of the participants after a close inspection of valid and invalid participants. We did not find other participant specific characteristics. Both subgroup of participants shared similar mean application session duration (p<0.03, with one-way ANOVA) with overlapping standard deviation (valid: 97 ± 421[s], invalid: 109 ± 475[s]). We extended our analysis based on those findings. We gathered all the application sessions from S2's participants and filtered out the applications not in the category presented by expected QoE (kept 41585 sessions from a total of 63529 sessions). Then, we grouped the sessions by periods: T0, T2, and T3. Interestingly, we found that 64% of the participants decreased the duration of their application session during T2 for all the application categories (p<0.02, decrease of -1.2 ± 178[s]) compared to T0 ($59.3 \pm 121[s]$). Additionally, only the communication and social category applications were used less in T3 than T0. However we found that the other categories of application were used more (p< 0.001, increase of $+39.2 \pm 190$[s]). Finally, we observed an increase in application session duration for 65% of the participants (p<0.01, increase of $+40.5 \pm 217$[s]) between T2 and T3. The increase is an unexpected outcome. However, it could be explained by the participant's fatigue in the study or the random QoE level shown during T3, negatively influencing their attitude despite expectQoE.

4.9 ExpectQoE Model: QoE Levels and Features (S2)

The expectQoE model achieved high accuracy for the majority of the S2 participants (i.e., Section 4.7). Hence, we quantified the QoE of all the application usage sessions (63529 sessions in S2) using the features previously selected (Table 1). Table 2 shows the QoE level per participants in T0. We found that overall in S2 the majority of the session was of high QoE level (86% ± 7) and a minority of low QoE (13% ± 7).

In order to better understand our results, the statistical significance was derived using a one-way ANOVA test, as both high and low median QoE level follow a normal distribution (high: p<0.002, low: p< 0.008). Our analysis revealed that the median application duration was lower in low QoE sessions, 26±9 seconds, contrary to 29±10 seconds for high QoE sessions (p<0.001). Hence, the participants spend more time in sessions rated as high QoE across all S2 sessions.

Furthermore, we explored the impact of the participant's physical activity and network state on the QoE level. The Radio Access Technology (RAT, e.g., Wi-Fi, LTE, UMTS, ...) does not influence the QoE level, and twenty-four participants had the same top RAT distribution for high and low QoE sessions. Then, we focused on the cell signal strength and the Wi-Fi signal (dBm). There were no significant differences between high and low QoE sessions based on cell signal strength values (dBm). Further analysis showed that the median Wi-Fi signal strength was lower in low QoE session -72±3 dBm (weak, one bar on screen) than in high QoE session -65±3 dBm (fair, two bars on screen) (p<0.01). Finally, on average, the participants obtained a high QoE when their physical activity was "still" (64%±19) and lower on the other activities (36%±16) like walking. However, these results were not significant (p>0.1).

5 DISCUSSION AND LIMITATIONS

In summary, our research aimed to explore the influence of a QoE level notification system (expectQoE) on smartphone application users. We verified expectQoE influence through qualitative and quantitative data. The second aim was to ascertain whether the factors influencing the QoE of smartphones have changed in the last decade. First, the analysis of the real model QoE predictions against the participants' reported QoE level yielded significant results: expectQoE influenced the participants' application usage duration. However, a the relatively low performance was reported by the participants, which is different from the performance obtained during the model building phase. It could be caused by several aspects: limited training data which incorporate all the possible features' combination; bias caused by model generalization; model structure and parameter tuning. A personalized participant model could be another approach to enhance the model performance. The model would be fine-tuned for each participant via reinforcement learning or hyperparameters optimization. The resulting models would embody one's way of perceiving the level of QoE. The user's intent may play a role in this context, where the utilitarian needs to satisfy their intent is ranked higher than their needs of hedonic satisfaction [20] and the knowledge that QoE is going to be low.

Second, we found a statistical difference in the system perception on the participants between the real model predictions (T2) and than the random model (T3), validating the model performance. Third, the MCCM analysis found that the expectQoE notifications drive their application usage duration. Also, the effect size is stronger when the notifications contain low QoE indications. On one hand, this could be explained by smartphone's users preemptively limiting the time they spend in an application to reduce their predicted annoyance. On the other hand, the MCCM results are difficult to generalize (from only 17 valid participants over 30,Table 3).

Four, we found a significant trend in the application usage duration once the intervention started (i.e., decrease application duration). However, this effect is unsustainable in time due the participant fatigue in the study or the impact of the random notification (T3), decreasing their trust in expectQoE. Additionally, the QoE level was high overall. Hence, the need for expectQoE interactions may be only suitable and useful for specific contexts (e.g., physical activity changes, roaming, and optimizing smartphone use duration to satisfy the user's intent faster and reduce their smartphone usage).

Fifth, the factors influencing the QoE of popular smartphone applications remain unchanged since documented in 2012. However, we found that smartphone users are more network conscious and care about the impact of their smartphone usage on their digital wellbeing. Also, they subscribe to multiple streaming services and often have unlimited internet access (no data cap). These changes can be linked to the smartphone entering the plateau of productivity (i.e., mainstream adoption) [17], in contrast to 2012 [18], in which smartphones were on the rise of adoption.

Finally, we expect the implications of our work can help smartphone application developers to enhance their software by going beyond simple network state indications, and include a QoE aware system, which is capable to preemptively notify to their users an approaching low QoE event. Application developers could learn from our work by making context an intrinsic information source in their application performance evaluation. Hence, enabling them to build a better experience metric than scroll jank (e.g., visual hiccup and artifact) and startup latency.

Most limitations in the work presented in this paper arise from our choice to collect data in the wild with limited disturbance to our participants. Moreover, the dataset gathered in S1 and used for S2 modeling efforts was limited by the application usage collected and rated, mostly communication applications. We believe that in our case, the impact is limited due to communication being the most used application category. Nonetheless, the categories may not be sufficient as such applications contain services that depend on distinct network models, e.g. a video chat and a voice call have different network needs [38]. Hence, the category may be good for a small-scale study focusing on popular applications used for the main service (e.g., WhatsApp for text message and not for video conference). We believe that user's intent within the application should be the focus in future studies. As well, the root cause of low QoE events should be explored. Although we found an impact of expectQoE on our participants' application use duration, the interaction model we used (dyad of emoji with randomized placement) could have influenced the participants if they expected to observe the categories' emojis in the same place. Furthermore, the applications observed in S1 were all internet-enabled. Thus, the model implemented in S2 focuses on this type of application. However, the method to collect data, build and deploy a QoE model presented could be applied to the non-internet application. The network quality indicator (bars) are insufficient for the user to assess its expected QoE level for offline application due to the multiple factors influencing their experience. Finally the timing of S1 and S2; S1 happened in 2018 and S2 in 2021. During these three years, the Android system evolved. The system upgrades could have impacted our results. However, the habits of the S2 participants were always compared with the data gathered in T0 (baseline usage). Hence, our overall findings are valid and the impact of this time difference is limited.

6 CONCLUSIONS

Through a mixed method of qualitative and quantitative data collection in which the participants were active in the research by providing information directly and indirectly, we presented our research regarding the effectiveness of a QoE-based notification system to limit smartphone users' burden in case of low QoE. First, this required gathering application usage QoE levels in situ (S1). Second, it required building a QoE classifier from the data obtained during S1. This classifier was then included in our smartphone logger, providing the participants in S2 with the expected QoEs through notification. Our results showed that the participants reported higher satisfaction when expectQoE showed the real model predictions (T2) rather than random QoE level (T3). However, a global model have limited prediction capabilities. Hence, our future work includes building dynamically personalized QoE model based on the user application usage and context habits. Additionally, we investigated whether the expectQoE notifications had an impact on participant application usage by employing a MCCM analysis on a time series constructed from different periods of S2. Overall, we found that expectQoE decreased application usage duration for some participants. The influence was stronger when low QoE notification was shown. We also identified some features (e.g., Wi-Fi strength and physical activity) that impacted the overall QoE level of the participants during S2. Third, we presented changes in the factors influencing the QoE of the smartphone. We found that all factors are still applicable. However, they have evolved with new smartphone usages (e.g., streaming audio and video content). Additionally, smartphone users are now more network and wellbeing conscious than ever.

ACKNOWLEDGMENTS

SNSF MIQmodel (157003) (2015-2019), *AAL Guardian* (2019-6-120-CP), Swissuniversities P-13 project and UNIGE COINF (2018-2020) . The University of Geneva ethics committee approved those studies under *CUREG.201803.02* and *CUREG,201909.12*.

REFERENCES

[1] Ozgu Alay, Vincenzo Mancuso, Anna Brunstrom, Stefan Alfredsson, Marco Mellia, Giacomo Bernini, and Hakon Lonsethagen. 2018. End to End 5G Measurements with MONROE: Challenges and Opportunities. *IEEE 4th International Forum on Research and Technologies for Society and Industry, RTSI 2018 - Proceedings* (2018). https://doi.org/10.1109/RTSI.2018.8548510 ISBN: 9781538662823.

[2] Vance W. Berger and YanYan Zhou. 2014. Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society. https://doi.org/10.1002/9781118445112.stat06558 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat06558.

[3] Allan Berrocal, Waldo Concepcion, Stefano De Dominicis, and Katarzyna Wac. 2020. Complementing Human Behavior Assessment by Leveraging Personal Ubiquitous Devices and Social Links: An Evaluation of the Peer-Ceived Momentary Assessment Method. *JMIR mHealth and uHealth* 8, 8 (2020), e15947. https://doi.org/10.2196/15947 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[4] Allan Berrocal, Vlad Manea, Alexandre De Masi, and Katarzyna Wac. 2020. mQoL Lab: Step-by-Step Creation of a Flexible Platform to Conduct Studies Using Interactive, Mobile, Wearable and Ubiquitous Devices. *Procedia Computer Science* 175 (Jan. 2020), 221–229. https://doi.org/10.1016/j.procs.2020.07.033

[5] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. 2011. Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 47–56. https://doi.org/10.1145/2037373.2037383

[6] Pedro Casas, Alessandro D'Alconzo, Florian Wamser, Michael Seufert, Bruno Gardlo, Anika Schwind, Phuoc Tran-Gia, and Raimund Schatz. 2017. Predicting QoE in cellular networks using machine learning and in-smartphone measurements. *2017 9th International Conference on Quality of Multimedia Experience, QoMEX 2017* 02152 (2017), 3–8. https://doi.org/10.1109/QoMEX.2017.7965687 ISBN: 9781538640241.

[7] Pedro Casas, Michael Seufert, Florian Wamser, Bruno Gardlo, Andreas Sackl, Raimund Schatz, and others. 2016. Next to You: Monitoring Quality of Experience in Cellular Networks from the End-devices. *IEEE Transactions on Network and Service Management* 4537, c (2016), 1–1. https://doi.org/10.1109/TNSM.2016.2537645

[8] Pedro Casas, Juan Vanerio, and Kensuke Fukuda. 2017. GML learning, a generic machine learning model for network measurements analysis. In *2017 13th International Conference on Network and Service Management (CNSM)*. 1–9. https://doi.org/10.23919/CNSM.2017.8255998 ISSN: 2165-963X.

[9] Pedro Casas, Martin Varela, Pierdomenico Fiadino, Mirko Schiavone, Helena Rivas, and Raimund Schatz. 2015. On the analysis of QoE in cellular networks: From subjective tests to large-scale traffic measurements. *IWCMC 2015 - 11th International Wireless Communications and Mobile Computing Conference* (2015), 37–42. https://doi.org/10.1109/IWCMC.2015.7289054 ISBN: 9781479953448.

[10] Qi Alfred Chen, Haokun Luo, Sanae Rosen, Z. Morley Mao, Karthik Iyer, Jie Hui, Kranthi Sontineni, and Kevin Lau. 2014. QoE Doctor : Diagnosing Mobile App QoE with Automated UI Control and Cross-layer Analysis. In *Proceedings of the 2014 Conference on Internet Measurement Conference - IMC '14*. ACM Press, New York, New York, USA, 151–164. https://doi.org/10.1145/2663716.2663726

[11] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[12] Alexandre De Masi and Katarzyna Wac. 2018. You're Using This App for What? A mQoL Living Lab Study. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 612–617. https://doi.org/10.1145/3267305.3267544

[13] A De Masi and K. Wac. 2019. Predicting quality of experience of popular mobile applications from a living lab study. In *2019 11th International Conference on Quality of Multimedia Experience, QoMEX 2019*. Berlin. https://doi.org/10.1109/QoMEX.2019.8743306

[14] Alexandre De Masi and Katarzyna Wac. 2020. Towards accurate models for predicting smartphone applications' QoE with data from a living lab study. *Quality and User Experience* 5, 1 (Oct. 2020), 10. https://doi.org/10.1007/s41233-020-00039-w

[15] Alan Dix. 2020. Statistics for HCI: Making Sense of Quantitative Data. *Synthesis Lectures on Human-Centered Informatics* 13, 2 (April 2020), 1–181. https://doi.org/10.2200/S00974ED1V01Y201912HCI044

[16] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia. 2010. A generic quantitative relationship between Quality of Experience and Quality of Service. *Blekinge Tekniska hogskola* 24, March (2010), 36–41. https://doi.org/10.1109/MNET.2010.5430142 ISBN: 0890-8044 VO - 24.

[17] Gartner. 2021. Gartner Hype Cycle Research Methodology. https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

[18] S Ickin, K Wac, M Fiedler, L Janowski, Hong Jin-Hyuk, and a K Dey. 2012. Factors influencing quality of experience of commonly used mobile applications. *Communications Magazine, IEEE* 50, April (2012), 48–56. https://doi.org/10.1109/MCOM.2012.6178833 ISBN: 0163-6804.

[19] ITU-T Recommendation P.800.1. 2019. Mean Opinion Score (MOS) Terminology. (2019), 1–8.

[20] Daniel Kahneman, Ed Diener, and Norbert Schwarz (Eds.). 1999. *Well-Being: Foundations of Hedonic Psychology*. Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610443258

[21] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. 2004. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* 306, 5702 (Dec. 2004), 1776–1780. https://doi.org/10.1126/science.1103572 Publisher: American Association for the Advancement of Science Section: Report.

[22] Fabian Kaup, Florian Fischer, and David Hausheer. 2017. Measuring and predicting cellular network quality on trains. In *2017 International Conference on Networked Systems (NetSys)*. 1–8. https://doi.org/10.1109/NetSys.2017.7903960

[23] Patrick Le Callet, Sebastian Möller, and Perkis Andrew. 2012. Qualinet White Paper on Definitions of Quality of Experience. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)* March (2012).

[24] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 770–780. https://doi.org/10.1145/2971648.2971724

[25] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-Aware Computing: Modelling User Engagemet from Mobile Contexts. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16* (2016), 622–633. https://doi.org/10.1145/2971648.2971760 ISBN: 9781450344616.

[26] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1021–1032. https://doi.org/10.1145/2858036.2858566

[27] Alberto Monge Roffarello and Luigi De Russis. 2019. The Race Towards Digital Wellbeing: Issues and Opportunities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300616

[28] Leanne G. Morrison, Charlie Hargood, Veljko Pejovic, Adam W. A. Geraghty, Scott Lloyd, Natalie Goodman, Danius T. Michaelides, Anna Weston, Mirco Musolesi, Mark J. Weal, and Lucy Yardley. 2017. The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone-Based Stress Management Intervention: An Exploratory Trial. *PLOS ONE* 12, 1 (Jan. 2017), e0169162. https://doi.org/10.1371/journal.pone.0169162 Publisher: Public Library of Science.

[29] Adrià Muntaner-Mas, Victor A Sanchez-Azanza, Francisco B Ortega, Josep Vidal-Conti, Pere Antoni Borràs, Jaume Cantallops, and Pere Palou. 2021. The effects of a physical activity intervention based on a fatness and fitness smartphone app for University students. *Health Informatics Journal* 27, 1 (Jan. 2021), 1460458220987275. https://doi.org/10.1177/1460458220987275 Publisher: SAGE Publications Ltd.

[30] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 91:1–91:25. https://doi.org/10.1145/3130956

[31] Raimund Schatz and Sebastian Egger. 2011. Vienna surfing : assessing mobile broadband quality in the field. In *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack - W-MUST '11*. ACM Press, New York, New York, USA, 19. https://doi.org/10.1145/2018602.2018608

[32] Raimund Schatz, Sebastian Egger, and Alexander Platzer. 2011. Poor, good enough or even better? Bridging the gap between acceptability and QoE of mobile broadband data services. *IEEE International Conference on Communications* May 2014 (2011), 6. https://doi.org/10.1109/icc.2011.5963220 ISBN: 9781612842332.

[33] Anika Schwind, Cise Midoglu, Ozgu Alay, Carsten Griwodz, and Florian Wamser. 2020. Dissecting the performance of YouTube video streaming in mobile networks. *International Journal of Network Management* 30, 3 (2020), e2058. https://doi.org/10.1002/nem.2058 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nem.2058.

[34] Arthur A Stone and Saul Shiffman. 1994. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine* 16, 3 (1994), 199–202. https://doi.org/10.1093/abm/16.3.199 Publisher: Lawrence Erlbaum Place: US.

[35] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. 2012. Detecting Causality in Complex Ecosystems. *Science* 338, 6106 (Oct. 2012), 496–500. https://doi.org/10.1126/science.1227079 Publisher: American Association for the Advancement of Science.

[36] Channary Tauch and Eiman Kanjo. 2016. The roles of emojis in mobile phone notifications. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1560–1565. https://doi.org/10.1145/2968219.2968549

[37] International Telecommunication Union. [n.d.]. E.800 : Definitions of terms related to quality of service. https://www.itu.int/rec/T-REC-E.800-200809-I

[38] Dimitris Tsolkas, Eirini Liotou, Nikos Passas, and Lazaros Merakos. 2017. A survey on parametric QoE estimation for popular services. *Journal of Network and Computer Applications* 77, October 2016 (2017), 1–17. https://doi.org/10.1016/j.jnca.2016.10.016 Publisher: Elsevier.

[39] Niels van Berkel, Simon Dennis, Michael Zyphur, Jinjing Li, Andrew Heathcote, and Vassilis Kostakos. 2020. Modeling interaction as a complex system. *Human–Computer Interaction* 36, 2021 (Jan. 2020), 1–27. https://doi.org/10.1080/07370024.2020.1715221

[40] Mariek M P Vanden Abeele. 2021. Digital Wellbeing as a Dynamic Construct. *Communication Theory* 31, 4 (Nov. 2021), 932–955. https://doi.org/10.1093/ct/qtaa024

[41] Hao Ye, Richard J. Beamish, Sarah M. Glaser, Sue C. H. Grant, Chih-hao Hsieh, Laura J. Richards, Jon T. Schnute, and George Sugihara. 2015. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences* 112, 13 (March 2015), E1569–E1576. https://doi.org/10.1073/pnas.1417063112 Publisher: National Academy of Sciences Section: PNAS Plus.