Efficacy of Mobile Context-aware Notification Management Systems: A Systematic Literature Review and Meta-Analysis

Florian Künzler
Department of Management,
Technology and Economics
ETH Zurich
Zurich, Switzerland
Email: fkuenzler@ethz.ch

Jan-Niklas Kramer
Institute of Technology Management
University of St. Gallen
St. Gallen, Switzerland
Email: jan-niklas.kramer@unisg.ch

Tobias Kowatsch
Institute of Technology Management
University of St. Gallen
St. Gallen, Switzerland
Email: tobias.kowatsch@unisg.ch

***Abstract*—Notifications can be relevant but they can also decrease productivity when delivered at the wrong point in time. Smartphones are increasingly capable of detecting relevant context information with the goal to decrease the number of these badly timed interruptions. Accordingly, research on context-aware notification management systems (CNMSs) on mobile devices has received increasing attention recently, prototypes have been built and empirically evaluated. However, there exists no systematic overview of mobile CNMSs evaluating their efficacy. The objectives of the current work are therefore to identify relevant empirical studies that have assessed the efficacy of mobile CNMSs and to discuss the findings with respect to future work. A systematic literature review and meta-analysis was conducted to address these objectives. Consistent with prior work, two efficacy metrics were applied: response rate and response delay. A keyword-based search strategy was used and resulted in 1'634 studies, out of which 8 were relevant for the topic. Findings indicate that mobile CNMSs increase the response rate, while there was only little evidence that they reduce response time, too. Implications for researchers and practitioners are discussed and future research is outlined that aims at further increasing the efficacy of mobile CNMSs.**

I. Introduction

The number of mobile devices is growing tremendously [1], [2], and so is the number of notifications these devices deliver to their users. In 2015, an individual received 100 notifications on average every day [3]. Notifications interrupt individuals and decrease task performance [4], [5]. Consequently, interruptions are associated with costs and are also associated with frustration [6] and an individual's feeling of being constantly interrupted by his or her computer system [7]. While the number of notifications has increased, the attention of an individual has remained constant.

Context-aware notification management systems (CNMSs) have received increasing attention as a potential solution for this problem. These systems infer the context (e.g. location or surrounding) of the user to make an informed decision on when to interrupt the user with a notification. Over the last decade first prototypes and empirical studies have been conducted [8], [9]. While Okoshi et al. [8] reviewed several papers and discussed CNMSs as a layer in-between notification reception and delivery to the user, Turner et al. [9] analyzed interruptibility (i.e. a users susceptibility to receive a notification), data collection and prediction scenarios. Most of the work, however, has been focusing on desktop notifications and using external sensors. Horwitz et al. [10] and Fogarty et al. [11] conducted some of the earliest work in this field and find that simple simulated sensors (i.e. door open / closed, talk, telephone usage etc.) provide enough information to infer interruptibility [11]. Lilsys [12] and BusyBody [13] were more advanced systems implementing on-the-fly detection of interruptibility using third party sensors and computers, respectively. The sensors they used range from motion, sound (both external and stationary sensors) and time to location, meeting status (in a meeting or not) and computer activity (e.g. switching of applications). Subsequently, Iqbal et al. [14], [15] developed OASIS, a system using application switching on desktop computers to trigger notifications. All these systems focus on detecting naturally occurring breakpoints (i.e. breakpoints in the daily schedule of a user, e.g. ending a phone call), by either sensing task breakpoints (i.e. a break within a task) or task boundaries (i.e. the start and stop of a task) as opportune times to send out notifications. With the increasing adoption of mobile devices, desktop computers were not the only source of interruptions anymore, therefore mobile CNMS received increasing attention. The first step from desktop computers to mobile devices are body worn sensor networks, investigated by Ho et al. [16]. They find that notifications delivered at activity breakpoints are better received. Modern mobile devices, with smartphones in particular, provide a unique set of sensors. These sensors can be used to infer the context of the user to time notifications more informed, as the device is mostly very close to the user. Notable work comes for example from Mathur et al. [17], who studied smartphone interruptibility with EEG sensors to extract which sensor information from the smartphone is relevant. They then trained different algorithms and machine learning models and found that personalized models are only marginally better than models trained on a large population.

However and to the best of our knowledge, there exists no systematic overview of CNMSs that has evaluated the efficacy of CNMS on mobile devices using only device sensors. The objectives of the current work are therefore (1) to identify relevant empirical studies that have assessed the efficacy of mobile CNMS and (2) to discuss the findings with respect to future work. In order to address these goals, we conducted a systematic literature review and meta-analysis. Consistent with prior work [18], [19], several efficacy metrics were applied. Particularly, response rate, i.e. the number of notifications a user has processed divided by the total number of notifications sent out [18], and response delay, i.e. the time between a notification has been sent and was processed by the user [19], were used as objective efficacy metrics.

Next, we describe our systematic literature review including the search strategy, study selection procedure, and details about the meta analysis and risk bias assessment. Then, the results are presented and discussed. We finally outline future research opportunities that should be taken into account to further increase the efficacy of mobile CNMS.

II. Method

We followed the established guidelines for systematic literature reviews as outlined by Okoli et al. [20] and Webster et al. [21] to make the search process of the current work transparent. We first outline how we defined the keywords and found the relevant databases. Thereafter, we describe how the relevant articles were identified. Finally, we describe the methods that were applied to conduct our results analyses.

A. Search strategy

We first conducted a preliminary search on Google Scholar for articles related to interruptibility. This unstructured search resulted in a first overview of the field of interruptibility and we learned that the papers of interest were predominantly technical in the field of CNMSs with some papers having a health (e.g. [22], [19]) or psychological (e.g. [23]) background. Consistent with prior work [24], we used the following databases to conduct our search: ACM Digital Library, IEEE Explore, ScienceDirect and ProQuest. To cover health related articles we included the PubMed database. PubMed provides access to MEDLINE with 16 million references one of the largest and most widely used database in health related research [25]. We conducted our search in February 2017 and looked for papers ranging back to 1994, which is considered as the year the smartphone was invented [26]. Consistent with prior work [27] we decided to use a keyword based search strategy to efficiently screen a wide range of articles and journals. The keywords are depicted in Table I. All keywords were derived from papers we found in our early unstructured search. For an article to be included in the results at least one keyword from every set: *Title*, *Abstract 1* and *Abstract 2* had to be included (see Table I). To decide on the relevancy of an article, we screened the title and abstract for evidence regarding the application of a CNMS as described in the introduction.

After this pre-screening we applied the formalized inclusion criteria described in the next section. We then conducted a backward search on the papers passing this inclusion criteria catalogue and applied the same procedure to the references of the included papers.

TABLE I
Keyword combinations

Title (*OR*)	Abstract 1 (*OR*)	Abstract 2 (*OR*)
notification	smartphone	
interrupt	’smart phone’	notification
attention	smartwatch	interrupt
intelligent	’smart watch’	intelligent
’context-aware’	’mobile phone’	breakpoint
’state of receptivity’	’cellular phone’	’state of receptivity’
’state of vulnerability’	phone	’state of vulnerability’
’state of opportunity’	’mobile device’	’state of opportunity’
	’push notification’	

B. Study selection

Following Okoli et al. [20] we developed a list of inclusion criteria a paper has to fulfill to be considered as relevant. The purpose of these criteria is to ensure relevancy with respect to the research objectives. Before applying these inclusion criteria to the identified studies we discussed them among the authors to ensure a common understanding. We then applied the criteria independently to the identified papers and discussed any disagreement in the team of authors.

We included studies, which fulfilled the following criteria. First, all studies were screened for some quality criteria (1) *English papers only.* (2) *Only peer-reviewed papers.* (3) *Only RCTs and crossover designs.* To infer causality, we excluded observational studies and expert opinions. Second, all studies were screened for relevance to the review (4) *Only studies with notifications on mobile devices, i.e. no desktop notifications.* (5) *Only studies, who’s only source of data is coming from the mobile device are included, i.e. no third party devices or services.* (6) *Only studies evaluating a CNMS by automatically inferring an opportune timing to send out notifications*. (7) *Only studies, who reported efficacy metrics of the CNMS by response delay and / or response rate*. (8) *Only working CNMS prototypes and products.* While there are many data gathering and concept studies, we were interested in systems that have an algorithm in place to detect and exploit interruptibility.

C. Meta analysis

Response rate and response delay returned dichotomous data values (e.g. did the user respond or not) and continuous data (e.g. delay in minutes), respectively. To compare dichotomous data from different studies, we used the odds ratio (OR), as suggested by the *Cochrane Handbook for Systematic Reviews of Interventions* [28]. The OR was calculated with the software RevMan 5.3 [29]. The analysis was done using the Mantel-Haenszel method [30], [31] with random effects [28]. This method has been shown to have better statistical properties for a small number of studies [31], [28], which is the case in our analysis. The random effects option will give a more conservative result if heterogeneity is found, but will give the same result as fixed effects if the data shows no heterogeneity [28]. For continuous data we used Cohen's d as effect size, as suggested by [32]. To calculate d from the different sources of information in the articles we used *www.psychometrica.de* [33]. To compare the different studies against each other we used the approach described by Thiese et al. [34] to classify the study design.

D. Risk bias assessment

The risk of a potential bias in the publication was assessed depending on the study design. For randomized controlled trials (RCT) we assessed the risk as recommended by the *Cochrane Handbook for Systematic Reviews of Interventions* in Chapter 8 [35]. Specifically, their criteria catalogue can be applied if the study of interest has a selection bias (i.e. systematic differences between the groups, which are compared [36]), performance bias (i.e. systematic differences between groups with regard to exposure to other factors that might affect the outcome [36]), detection bias (i.e. differences between groups in how outcomes are assessed [36]) or attrition bias (i.e. bias due to the amount, nature or handling of missing outcome data [36]). For crossover studies, i.e. studies with multiple treatments per participant, we used Chapter 16 of the handbook by [36] and the CONSORT N-of-1 guidelines [37]. Specifically, we assessed the risk of a potential carry-over effect (i.e. persistence of effects of one treatment into a later period of treatment [37]), a period effect (i.e. a change of outcome over time even in the absence of treatment [37]), a sequence effect (i.e. systematic differences between groups with predefined treatment sequences), not accounting for non-independence of outcome data in analysis (i.e. not using some form of a paired analysis) and a potential attrition bias.

III. Results

In this section, we first report the general findings related to our search strategy and describe the study characteristics of the selected articles. We then provide an overview of the study designs and potential biases in the studies. Finally, we explain which sources of context information were used by the mobile CNMSs and, most importantly, we report the results of the efficacy metrics and their connection to the different sources of context information.

A. Search strategy results

Our search process revealed a total of 1'634 articles. Out of these, 135 articles passed the first title and abstract-based pre-screening step and 9 articles the second screening step in which the inclusion criteria were applied. During the backward search we analyzed the references of these 9 articles as well, but no further articles of interest were found. Furthermore, one article dropped out for duplicate reporting of a study result resulting in 8 relevant articles.

B. Study characteristics

Table II provides a summary of the study characteristics of all 8 articles of our systematic literature review. Age and gender distributions of the study participants indicate a trend towards males aged between 20 and 30. There are only two studies [19], [18] including older participants (i.e. with a mean age around 50) but four studies [7], [38], [39], [40] with participants younger than 30. Also, there is a gender imbalance, as only one study [41] established gender balance, while all other studies had more men than women with the gender ratio women to men going as low as 1 to 3 [7].

C. Risk of bias

Due to the lack of adequate reporting, risk of bias could not be completely assessed for all included studies. For example, we were unable to assess risk of bias in the included RCT study [19] because the authors did not report a description of the randomization process (i.e. how the randomization sequence was generated) nor a comparison of baseline characteristics between the different groups or how missing data was handled in the analysis. For crossover studies, the balance with regard to group assignment over time is essential to establish causal inference [36]. Most studies randomized participants to different notification strategies multiple times over the course of the study. Randomization, if properly implemented, results on average in a balanced design (i.e. a balanced order of notification strategies), which is necessary to avoid confounding of the effect of interest with the order of notification strategies (carry-over effect) or changes in the outcome that occur naturally over time (period effect). This works well when a large number of participants are randomized or the order of the notification strategies is altered often. Consequently, we judged the risk of a period effect to be high in three [6], [7], [39] out of seven studies. Only one study [38] implemented an *apriori* balanced study design. In addition, studies rarely reported how missing data was handled in the analysis and we thus were not able to assess attrition bias in four studies [7], [18], [19], [39]. Because the main outcome in most studies was a stable behavioral variable (e.g. response time), we assumed the risk for carry-over effects to be low in almost all studies. The detailed risk of bias assessment is reported in Table III. In two studies [40], [38] it was unclear which statistical analysis was used. All in all, the results of our analysis have to be interpreted with great caution since risk of bias in all included studies was judged to be high or unclear.

D. Efficacy metrics

Results with regard to the efficacy metrics response rate and response delay are described in the following subsections.

As outlined in Table II six studies (75%) reported results on the response rate. Four studies (50%) contained enough information to compute the odds ratio, as shown in Table IV. The four analyzed studies show that there is a statistically significant effect of mobile CNMSs improving the response rate to notifications. The weighted odds ratio is 1.71 with a 95% CI of [1.28, 2.28] and an I^2 of 75%, justifying the

TABLE II
SUMMARY OF ALL STUDIES IN THIS SYSTEMATIC REVIEW.

Paper	Participants			Study Design	Method Used	Response Rate	Response Delay
	N	Age	% Women			Efficacy Metrics	
Morrison 2017 [19]	77	18-62	49%	RCT	Context-based timing	x	x
Obuchi 2016 [38]	30	18-26	36%	Randomized crossover study	Breakpoint based timing	x	x
Okoshi 2016 [39]	30	18-29	37%	Randomized multiple crossover study	Context-based timing		x
Okoshi 2015 [7]	41	19-26	24%	Randomized multiple crossover study	Context-based timing	x	
Pielot 2015 [18]	16	16-51	N/A	Non-randomized multiple crossover study	Boredom-triggered timing	x	
Pejovic 2014 [40]	10	22-26	40%	Randomized multiple crossover study	Context-based timing		x
Fischer 2011 [41]	20	21-48	50%	Non-randomized multiple crossover study	Timed after phone call or text message	x	x
Fischer 2010 [6]	11	N/A	27%	Non-randomized multiple crossover study	User defined timing	x	

TABLE III
RISK OF BIAS FOR ALL STUDIES

Crossover	N	DR	L	CE	PE	SqE	AB	NiA
Fischer [6]	11	0%	10	+	+	N/A	-	y
Fischer [41]	20	0%	14	-	?	N/A	-	y
Pejovic [40]	10	0%	30	-	-	N/A	-	?
Okoshi [7]	41	0%	31	-	+	N/A	?	y
Pielot [18]	16	0%	12	-	-	N/A	?	y
Obuchi [38]	30	7%	4	-	-	-	-	?
Okoshi [39]	30	10%	16	-	+	N/A	?	y
RCT	N	DR	L	SB	PB	DB	AB	
Morrison [19]	77	0%	14	?	N/A	-	?	

y=yes, n=no, +=high, -=low, ?=unclear, N=Sample size, DR=Dropout rate, L=Length [days], CE=Carry-over effect, PE=Period effect, SqE=Sequence effect, AB=Attrition bias, NiA=Accounted for non-independence in analysis, SB=Selection bias, PB=Performance bias, DB=Detection bias

random effects model. According to Chen et al. [42] an odds ratio of 1.68, 3.47 and 6.71 correspond to a small, medium or large effect, respectively. Consequently, an odds ratio of 1.71 equals a small effect. The fifth study [18], which did not have enough data reported to be included in our calculation, supports our finding with $r = .51$, which is equal to an odds ratio of 8.64 [33], i.e. a large effect. The last study [6] reporting a response rate did not find a statistically significant effect ($\chi^2(1) = .004$, $p = 1.0$).

Three studies [39], [41], [40] reported sufficient information to calculate Cohen's d, while one study [38] reported a descriptive trend. Finally, [19] reported all necessary results, yet the direction of the effect was reported inconsistently, which is why this study was omitted in this analysis. Fischer et al. [41] achieved an effect size (Cohen's d) for the response delay, i.e. the difference in response delay between mobile CNMS and non-CNMS, of $d = .46$ ($F(2, 1374.9) = 73.71$, $p < .001$), Okoshi et al. [39] $d = .14$ ($Z = -3.19$, $p < .05$) and Pejovic et al. [40] $d = .14$ ($t(141.02) = 1.9$, $p = .06$). Cohen et al. [43] suggest a small effect for $.2 < d \leq .5$, a medium effect for $.5 < d \leq .8$ and a large effect for $d > .8$. Consequently, one study [41] has a small statistically significant effect, while the other two studies [39], [40] have a negligible effect. Obuchi et al. [38], who reported a descriptive result, found a trend towards a reduced response delay.

E. Sources of context information

A total of ten different sources of context information were used to evaluate the interruptibility of the user, as shown in Table V. Accelerometers were used to identify the movement pattern of the user, sometimes in combination with physical activity, a higher-level feature (e.g. to indicate whether a user has stopped walking which defines a natural breakpoint). Geographic location was usually inferred using GPS (e.g. to infer whether an individual has arrived at home or at work), however due to it's high power consumption was sometimes replaced with a less energy intense cell tower location method, at the cost of reduced accuracy. Microphones were mostly used to detect environment noise or people around the designated user (e.g. to identify whether a person was involved in social interactions). Bluetooth and WiFi were used to detect finer grain location changes (e.g. changing rooms). User interface (UI) events and communication relate to users' interaction with their phone. Where UI events were more related to interaction with the phone itself and communication more to interaction with other people through their phone, i.e. a call. Both of these features were used to detect naturally occurring breakpoints, such as the termination of a phone call or switching of applications on the smartphone. User defined rules is a special case for context awareness, as the user is asked beforehand to provide information about, when he or she can be interrupted with notifications. However, we decided to include it nonetheless, because the system later automatically decides on the interruptibility. Time and others refer to generally available information on the phone, such as battery status, time of the day, day of the week or whether

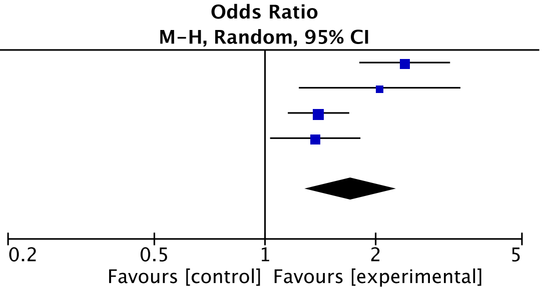
TABLE IV

ODDS RATIO FOR THE RESPONSE RATE. EVENT REFERS TO THE NOTIFICATIONS THAT WERE ANSWERED, TOTAL TO THE TOTAL NUMBER OF NOTIFICATIONS SENT. (SOURCE REVMAN [29])

	Experimental		Control			Odds Ratio
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**M-H, Random, 95% CI**
Okoshi 2015	290	364	647	1043	26.3%	2.40 [1.81, 3.19]
Morrison 2017	65	202	28	149	16.9%	2.05 [1.24, 3.40]
Fischer 2011	657	902	723	1100	30.5%	1.40 [1.15, 1.70]
Obuchi 2016	231	397	192	381	26.3%	1.37 [1.03, 1.82]
Total (95% CI)		**1865**		**2673**	**100.0%**	**1.71 [1.28, 2.28]**
Total events	1243		1590			

Heterogeneity: Tau^2 = 0.06; Chi^2 = 11.77, df = 3 (P = 0.008); I^2 = 75%

Test for overall effect: Z = 3.63 (P = 0.0003)



the phone is covered with something or not. These features are less about breakpoint detection, but more about general patterns of the users (i.e. no notifications during the night).

Table V summarizes the different sources of information used by the mobile CNMSs of the included studies and compares them with their efficacy. Some sources were used more commonly, such as location, accelerometer or time, while others are less common, like user defined breakpoints, microphone or Bluetooth.

Comparing the different sources with the respective performance in the efficacy metrics response rate and response delay gives a more detailed view on which sources seem to have a good predictive power to infer interruptibility. Using a single source of information, such as the accelerometer or communication pattern, provide a good starting point to improve the response rate [38] and [41]. Yet, using only user defined breakpoints did not prove to increase the response rate, as in [6]. However, adding more features, such as location or user interface events to the accelerometer features results in a further increased response rate, as [19] and [7] did, respectively. While we have too few data to compare communication patterns, we see that accelerometer based features can be improved with time and location or user interface events. Further increasing the number of different sources resulted in another increase of the response rate. While [18] did not include accelerometer data, they used WiFi, microphone, communication and other features as distinguishing sources and achieved the highest response rate.

Regarding response delay, we see that communication alone is the best feature to use to minimize delay [41]. Using more features without the communication feature results in worse performance, as seen in [39] and [40]. While, there is no information on the effect size for [38], they reported a descriptive trend towards a reduced response delay using only accelerometer features.

TABLE V

SOURCES OF INFORMATION USED TO INFER CONTEXT-AWARENESS IN THE DIFFERENT STUDIES

	Source of context information										Effect size	
	Location	Microphone	Bluetooth	UI Events	Communication	User defined	WiFi	Time	Accelerometer	Others	Response rate Odds ratio	Response delay Cohens d
[19]	x							x	x		2.1**	
[38]									x		1.4*	x^1
[39]				x								0.1*
[7]				x					x		2.4***	
[18]	x	x		x	x		x	x		x	8.6*	
[40]	x		x				x	x	x			0.1^{ns}
[41]					x						1.4***	0.5***
[6]						x					x^{ns}	

Others=additional phone information (e.g. battery status etc.), ^{1} found a descriptive trend towards a reduced response delay; significance levels: ns $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. p-values for the response rates based on [44].

IV. DISCUSSION

In this systematic review and meta analysis we identified eight relevant studies and assessed the efficacy of the corresponding mobile CNMSs with respect to response rate and response delay. We found that both gender and age of the participants were not balanced. In particular older people were less represented in the reviewed studies. For certain applications targeting a specific population (e.g. digital health interventions), this might be an issue. Also the small sample size in the investigated studies is a major concern, as the statistical power is reduced and consequently a true effect might be less likely [45]. A vast range of sensors and sources of data streams has been used to derive information about the context in which individuals are likely to be interruptible. In particular, sensor provided information (e.g. location based breakpoints) were superior to user provided static input (e.g. designated times for notifications). Consistently, Fischer et al. [6], who used user defined breakpoints, were not able to find an effect of mobile CNMSs on the response rate. It may be therefore not an easy task for individuals to predict their interruptibility. That is, while individuals might be correct about their general availability, it might be impossible for them to predict the exact point in time and context in which they will be interruptible.

The source of context information is a key component determining the efficacy of mobile CNMS. For response delay, no major effect could be observed, as for most of the investigated sensor combinations no effect was found. The major exception is the sole use of communication patterns, they showed the largest reduction of the response delay. This can be explained with the fact, that the smartphone is still in the hand of the user, but he or she stopped interacting with it. However, the authors of the paper note that, while it was a good way to reduce the response delay, the interruption for the user was seen as being very intrusive, since most of the time an immediate action is required after, for example, a phone call, i.e. taking notes. This also explains why the effect on the response rate was less impactful compared to other sources. Here, increasing the number of features increased the response rate. Accelerometer alone increased the response rate, yet the efficacy was improved, when more features, such as location or UI events were added. Adding UI events slightly outperformed the addition of time and location in increasing the response rate. Again because the user is interacting with the device and consequently more likely to see and immediately respond to a notification that is triggered. However, the best efficacy was achieved, when several different sources were used. The complexity of necessary insight into the users context serves as an explanation, why more features are needed to optimize for the response rate, than to optimize for the response delay.

With respect to the different sources of context information it is interesting to see what researchers have done, however, of equal interest are the sources we did not find. We did not find any study using the specific content of the notification influencing the context it was sent in. For example a nutrition related notification might be triggered before lunch, or while shopping, yet no study used the content of the notification as a source of information. Also none of the included studies used personality traits as a discriminating feature for the classification. There was one study [8] doing an analysis based on their gathered data and separating users into sensitive and insensitive users towards interruptibility (e.g. how positive or negative the feedback of individuals was towards an interruption). One study [40], which used online learning to personalize their models and Mathur et al. [17] training their models on the individual user's data. However, only Pejovic et al. [40] tested their system in the wild and no study has empirically evaluated further personalizing their systems using personality traits (i.e. the big five [46]). Instead of using the personality as a source of information, one could also use the inferred context to block all notifications in a certain environment, i.e. while driving. In particular, smartphones are capable of achieving this task, as they are mostly very close to the user. However, while we found some research outlining the concept of such systems [47], none of the included studies investigated this concept empirically. Also none of these studies touched on the privacy aspects of sharing and exploiting sensitive information for the purpose of sensing interruptibility.

With respect to the efficacy metrics, the current review shows that mobile CNMSs are able to increase the response rate compared to non-CNMS settings. Thus, a context-sensitive timing of notifications means that an individual is more likely to respond, lowering the probability that he or she forgets to respond. Against the background of our findings regarding response delay, i.e. we found only little evidence for mobile CNMSs reducing this efficacy metric, individuals probably need time to respond to but not to read a notification. Consequently, while the notification can theoretically be sent at any time and the user might see it very quickly, the response rate is negatively affected, if no mobile CNMS is used.

V. Limitations

No scientific work comes without limitations. We therefore outline three shortcomings in the following paragraph. First, we found a high risk or incomplete information for a bias assessment for all studies, which is why we are not able to interpret all efficacy metrics in a concise manner. Second, essential information missing in the articles reduced the number of studies we were able to include for the meta-analysis, further weakening our conclusions. And finally, not only the total number of studies, but also the number of participants in the individual studies were rather small and the gender and age distribution were biased. That is, the samples used in the studies were not representative with respect to a specific target population, which restricts our ability to find and discuss generalizable results.

VI. Implications and future research

Against the findings of the current work and the limitations described above, we now discuss distinct areas, where we see potential for future work that should be taken into account to further increase the efficacy of mobile CNMSs. First, a call for *mobile content and context-aware notification management systems (CCNMSs)*: As Okoshi et al. suggested in their literature review [8], most studies apply a layer structure, where the interruptibility mechanism is independent of the application. However, only few studies have analyzed what specific content means for interruptibility. We see great potential of specific targeted content in combination with a mobile CNMS on the response delay and response rate. For example, in the healthcare context, future research should test if adaptive interventions triggered by mobile CNMSs are more effective compared to a control condition with non-context-aware notifications. Our findings suggest an increased response rate. Consequently, we assume that the adherence to interventions is also higher, making health related interventions more efficient if mobile CNMSs are used.

Second, a call for *mobile personalized CNMSs*: While Mathura et al. [17], Okoshi et al. [39] and Pejovic et al. [40] did some first work on personalized CNMSs, we encourage more empirical research testing them in the field. Furthermore, we encourage research into mobile CNMSs applying reinforcement learning [48], [49], [50], as they provide great potential for self-improving systems adjusting to the special and dynamic circumstances of every user. Also behavioral questionnaires were not applied so far as a feature to personalize mobile CNMS (e.g. by evaluating the big five personality traits [51], [46]).

Third, a call for *safety optimized mobile CNMSs*: Our analysis has shown that different sources of context information can be used to infer interruptibility and since notifications can be a major source of distraction. We encourage research into using this context information for active safety optimized mobile CNMSs to prevent dangerous distractions in high concentration situation (e.g. while driving). While there are some concept studies of safety optimized mobile CNMS (e.g. [47]) and some efforts by large consumer electronics companies (e.g [52]), no study, to the best of our knowledge, has yet implemented and tested these systems empirically.

Fourth, a call for *open mobile CNMSs data*: Another opportunity for research on improving the efficacy of mobile CNMSs is to make sure that anonymized self-reports, behavioral data and usage log data collected by mobile CNMS is made freely available if individuals agree and provided there are no ethical, legal or copyright aspects negatively affected. For example, the Swiss National Science Foundation expects from October 2017 onwards that ”data generated by funded projects will be publicly accessible in non-commercial, digital databases” [53]. Consistently there exist already non-profit organizations, so called cooperatives in Switzerland, that provide corresponding data repositories, such as www.midata.coop or www.healthbank.coop. These services would allow researchers to identify individual (e.g. demographics, personality traits), behavioral (e.g. app usage characteristics), and contextual (e.g. social situations) patterns among others that could be used to further personalize the timing and content of mobile CNMSs and thus, their efficacy.

We finally make a call for *replication studies*: We would like to see future research testing the findings of the current work by considering standardized efficacy metrics of mobile CNMS in large-scale settings over an extended period of time with a representative sample, in the form of RCTs and with complete reporting. Only with this accumulated body of replication studies it will be possible to identify and better understand the driving factors that make mobile CNMSs effective.

VII. Conclusion

An increasing number of notifications is competing for the attention of individuals resulting in interruptions with serious consequences such as decreased task performance or stress. Context-aware notification management systems (CNMSs) have been therefore built to improve the timing of notifications. Against this background, the current work is the first that has evaluated the efficacy of mobile CNMSs with the help of a systematic literature review and meta analysis. Indeed, our findings indicate that mobile CNMSs are capable of increasing the response rate to notifications. However, we found only little evidence that response delay can be decreased, too. The most efficient source of context information to reduce the response delay was found to be communication patterns, as the user is still interacting with the phone and is interruptible. To increase the response rate, however, more complex combinations of information sources were required, with the study having the most sources, having the highest response rate. Moreover, the studies reviewed in this work have either a high risk of bias or do not contain the information required to draw general conclusions. We have therefore outlined several implications for future work to further increase the efficacy of mobile CNMSs.

VIII. Acknowledgement

We would like to thank David Kotz from Dartmouth College and the anonymous reviewers for their valuable suggestions that helped us to improve this paper.

References

[1] G. Inc. (2012, August) The new multi-screen world. [Online]. Available: https://www.thinkwithgoogle.com/research-studies/the-new-multi-screen-world-study.html

[2] Salesforce. (2014). [Online]. Available: https://www.marketingcloud.com/resource-center/digital-marketing/2014-mobile-behavior-report/

[3] A. Mehrotra, M. Musolesi, R. Hendley, and V. Pejovic, “Designing content-driven intelligent notification mechanisms for mobile applications,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 813–824.

[4] S. T. Iqbal and E. Horvitz, “Disruption and recovery of computing tasks: field study, analysis, and directions,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 677–686.

[5] D. McFarlane, “Comparison of four primary methods for coordinating the interruption of people in human-computer interaction,” *Human-Computer Interaction*, vol. 17, no. 1, pp. 63–139, 2002.

[6] J. E. Fischer, N. Yee, V. Bellotti, N. Good, S. Benford, and C. Greenhalgh, “Effects of content and time of delivery on receptivity to mobile interruptions,” in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. ACM, 2010, pp. 103–112.

[7] T. Okoshi, J. Ramos, H. Nozaki, J. Nakazawa, A. K. Dey, and H. Tokuda, “Reducing users’ perceived mental effort due to interruptive notifications in multi-device mobile environments,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 475–486.

[8] T. Okoshi, J. Nakazawa, and H. Tokuda, “Interruptibility research: opportunities for future flourishment,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1524–1529.

[9] L. D. Turner, S. M. Allen, and R. M. Whitaker, “Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 801–812.

[10] E. Horvitz and J. Apacible, “Learning and reasoning about interruption,” in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 20–27.

[11] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, “Predicting human interruptibility with sensors,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 1, pp. 119–146, 2005.

[12] J. B. Begole, N. E. Matsakis, and J. C. Tang, “Lilsys: sensing unavailability,” in *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 2004, pp. 511–514.

[13] E. Horvitz, P. Koch, and J. Apacible, "Busybody: creating and fielding personalized models of the cost of interruption," in *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 2004, pp. 507–510.

[14] S. T. Iqbal and B. P. Bailey, "Effects of intelligent notification management on users and their tasks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 93–102.

[15] ——, "Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 17, no. 4, p. 15, 2010.

[16] J. Ho and S. S. Intille, "Using context-aware computing to reduce the perceived burden of interruptions from mobile devices," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 909–918.

[17] A. Mathur, N. D. Lane, and F. Kawsar, "Engagement-aware computing: Modelling user engagement from mobile contexts," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 622–633.

[18] M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver, "When attention is not scarce-detecting boredom from mobile phone usage," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2015, pp. 825–836.

[19] L. G. Morrison, C. Hargood, V. Pejovic, A. W. Geraghty, S. Lloyd, N. Goodman, D. T. Michaelides, A. Weston, M. Musolesi, M. J. Weal *et al.*, "The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: An exploratory trial," *PloS one*, vol. 12, no. 1, p. e0169162, 2017.

[20] C. Okoli and K. Schabram, "A guide to conducting a systematic literature review of information systems research," *Sprouts Work. Pap. Inf. Syst*, vol. 10, no. 26, 2010.

[21] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly*, vol. 26, pp. xiii–xxiii, 2002.

[22] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support," *Annals of Behavioral Medicine*, pp. 1–17, 2016.

[23] I. Nahum-Shani, E. B. Hekler, and D. Spruijt-Metz, "Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework." *Health Psychology*, vol. 34, no. S, p. 1209, 2015.

[24] Y. Levy and T. J. Ellis, "A systems approach to conduct an effective literature review in support of information systems research," *Informing Science: International Journal of an Emerging Transdiscipline*, vol. 9, no. 1, pp. 181–212, 2006.

[25] *Key Healthcare databases*, King's College London, September 2014.

[26] I. Sager. (2012) Before iphone and android came simon, the first smartphone. [Online]. Available: https://www.bloomberg.com/news/articles/2012-06-29/before-iphone-and-android-came-simon-the-first-smartphone

[27] F. Wahle, L. Bollhalder, T. Kowatsch, and E. Fleisch, "Towards the design of evidence-based mental health information systems for people with depression: A systematic literature review and meta-analysis," *J Med Internet Res*, 2017.

[28] A. D. e. Deeks JJ, Higgins JPT, *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2011, vol. 4, ch. 9: Analysing data and undertaking meta-analyses.

[29] T. C. C. Copenhagen: The Nordic Cochrane Centre, "Review manager (revman) [computer program]. version 5.3," 2014.

[30] N. Manter and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *J Nat C Inst*, vol. 22, pp. 719–748, 1959.

[31] S. Greenland and J. M. Robins, "Estimation of a common effect parameter from sparse follow-up data," *Biometrics*, vol. 41, pp. 55–68, 1985.

[32] M. Borenstein, L. V. Hedges, J. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Wiley Online Library, 2009.

[33] W. Lenhard and A. Lenhard. (2016) Berechnung von effektstärken. [Online]. Available: https://www.psychometrica.de/effektstaerke.html

[34] M. S. Thiese, "Observational and interventional study design types; an overview," *Biochemia medica*, vol. 24, no. 2, pp. 199–210, 2014.

[35] S. J. e. Higgins JPT, Altman DG, *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2011, vol. 4, ch. 8: Assessing risk of bias in included studies.

[36] A. D. e. Higgins JPT, Deeks JJ, *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2008, ch. 16: Special topics in statistics.

[37] S. Vohra, L. Shamseer, M. Sampson, C. Bukutu, C. H. Schmid, R. Tate, J. Nikles, D. R. Zucker, R. Kravitz, G. Guyatt *et al.*, "Consort extension for reporting n-of-1 trials (cent) 2015 statement," *Journal of clinical epidemiology*, vol. 76, pp. 9–17, 2016.

[38] M. Obuchi, W. Sasaki, T. Okoshi, J. Nakazawa, and H. Tokuda, "Investigating interruptibility at activity breakpoints using smartphone activity recognition api," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1602–1607.

[39] T. Okoshi, H. Nozaki, J. Nakazawa, H. Tokuda, J. Ramos, and A. K. Dey, "Towards attention-aware adaptive notification on smart phones," *Pervasive and Mobile Computing*, vol. 26, pp. 17–34, 2016.

[40] V. Pejovic and M. Musolesi, "Interruptme: designing intelligent prompting mechanisms for pervasive applications," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014, pp. 897–908.

[41] J. E. Fischer, C. Greenhalgh, and S. Benford, "Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 2011, pp. 181–190.

[42] H. Chen, P. Cohen, and S. Chen, "How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies," *Communications in Statistics—Simulation and Computation*, vol. 39, no. 4, pp. 860–864, 2010.

[43] J. Cohen, "A power primer," *Psychological bulletin*, vol. 112, no. 1, p. 155, 1992.

[44] B. MedCalc Software, Ostend. (1993) Medcalc. [Online]. Available: https://www.medcalc.org/calc/odds_ratio.php

[45] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.

[46] L. R. Goldberg, "An alternative description of personality: the big-five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[47] J. Lindqvist and J. Hong, "Undistracted driving: a mobile phone that doesn't distract," in *Proceedings of the 12th workshop on mobile computing systems and applications*. ACM, 2011, pp. 70–75.

[48] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[49] B.-T. Zhang and Y.-W. Seo, "Personalized web-document filtering using reinforcement learning," *Applied Artificial Intelligence*, vol. 15, no. 7, pp. 665–685, 2001.

[50] Y.-W. Seo and B.-T. Zhang, "A reinforcement learning agent for personalized information filtering," in *Proceedings of the 5th international conference on Intelligent user interfaces*. ACM, 2000, pp. 248–251.

[51] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[52] A. Inc. (2017) Do not disturb while driving. [Online]. Available: https://www.apple.com/newsroom/2017/06/ios-11-brings-new-features-to-iphone-and-ipad-this-fall/

[53] H. U. Zürich. (2017) Swiss national science foundation introduces data management plans. [Online]. Available: www.oai.uzh.ch/en/news/553-2017-03-07-10-24-27