

Multiple Factors Mental Load Evaluation on Smartphone User Interface

Li, Meng; Albayrak, Armagan; Zhang, Yu; van Eijk, Daan

DOI

[10.1007/978-3-319-96059-3_33](https://doi.org/10.1007/978-3-319-96059-3_33)

Publication date

2018

Document Version

Final published version

Published in

Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)

Citation (APA)

Li, M., Albayrak, A., Zhang, Y., & van Eijk, D. (2018). Multiple Factors Mental Load Evaluation on Smartphone User Interface. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Auditory and Vocal Ergonomics, Visual Ergonomics, Psychophysiology in Ergonomics, Ergonomics in Advanced Imaging* (Vol. X, pp. 302-315). (Part of the Advances in Intelligent Systems and Computing book series; Vol. 827). Springer. https://doi.org/10.1007/978-3-319-96059-3_33

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright



Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



Multiple Factors Mental Load Evaluation on Smartphone User Interface

Meng Li^{1,2}(✉) , Armagan Albayrak¹ , Yu Zhang²,
and Daan van Eijk¹

¹ Delft University of Technology,
Landbergstraat 15, 2628CE Delft, Netherlands
m.li-4@tudelft.nl

² Xi'an Jiaotong University, Xianning Road 28, Xi'an 710029, China

Abstract. Smartphone is nowadays the most prevalent computer system, thus a lot of attention from academia and industries has been put to evaluate its quality of use. However, Smartphone has more complex interaction modes and usage scenarios than PC and laptop. And therefore assessing its quality using a conventional usability evaluation is not sufficient. Meanwhile, the mental load serves as an acknowledged index of effort that operators have put in human-machine interaction, especially under high-demanding context. Mental load contains a set of parameters in multiple dimensions, such as primitive task performance, biological measurement(s) and subjective mental load scale, which assesses the efforts of tasks under a particular environment and operating conditions. Thus, it is suitable for evaluating complex mental work, and may indicate the use of Smartphones.

The aim of this paper is to apply a multi-dimensional method to assess the mental load of users, and find out which measurement(s) is the most suitable one to evaluate the efforts for using a smartphone. During this study, the effort on conducting tasks with four difficulty levels were assessed using measurements in three dimensions, which were (1) user performance (task accomplishment and secondary task), (2) subjective rating (NASA-TLX scale) and (3) physiological function (EDA). The values of these measurements were compared across novice, average and skilled users. The results show that: task duration and number of usability error are significantly related with mental load and change with the difficulty level of tasks; in subjective rating, *Mental Demand*, *Effort* and *Frustration* were highly related with mental load.

Keywords: Mental load evaluation · Usability · Smartphone

1 Introduction

1.1 Quality of Use

Usability is an international standard for evaluating quality of use of computer systems, which is widely applied on vertical display terminals (VDTs). The first international standard mentioned usability is ISO/IEC 9126, which described “usability” as an index for assessing software quality from users’ perspective, and it should include

understandability, learnability, operability and attractiveness [1]. The acknowledged definition of “usability” is from ISO 9241, which defines usability as “The effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments” [2]. Though ISO 9241 did not regulate a uniform test method of these usability parameters, but it suggested the number of usability errors and task duration as variables for effectiveness and efficiency of a computer system. Li found that in real environment user’s behaviours often not conform to the action phases of Robicon model from motivational psychology. Thus, he suggested a compound user model [3], which is composed of cognitive and action errors that users make when performing a task, to evaluate the quality of use. However, the prevalence of Smartphone gave computational terminals more mobility and flexibilities under variable scenarios, while the conventional user testing mainly focus on static and single task setting, thus makes the usability evaluation of them more difficult.

According to ISO 9241-11, the cognitive demand in human-computer dialogue, also known as Mental Load, influences the usability, and therefore suggests a mental load evaluation method [2]. However, mental load measurement is not widely adopted on user interface evaluation. It is probably because of the complexity of mental workload measurement [4]. Mental load (ML) originates from the 1960s to evaluate complicated operation system of aircraft in a high-speed environment. In the 1980s, an integrated system of methodology on mental load measurement started establishing [5]. ML measured the efforts of operators when they execute tasks in specific environments and operational affordances. The researches on ML mainly focus on operation of aviation tasks and vehicle driving, in order to assess usability of man-machine interaction in these systems [6, 7]. Since the Smartphone currently integrated more and more multi-task functions, and often used under dynamic environment, the mental effort of the usage of a Smartphone could be comparable with the dual-task diagram on driving or flying a plane.

1.2 Mental Load Measurements

Because of the complexity of ML, in last three decades many measurements for ML were developed, which could be classified into three dimensions [8]:

Behaviour Measurements. They assess the behavioural performance of the operators, to estimate operators’ mental capacity objectively. Primitive task performance is often solely applied, or combined with secondary task(s) to measure the entire mental consumption of users [13, 14].

Subjective Measurements. They consist of structured or non-structured questions to probe ML perceived by the operators. Self-report, Cooper-Harper Questionnaire, NASA-TLX Questionnaire, SWAT Questionnaire, MRQ are widely applied methods for subjective measurement [9–12]. NASA-TLX proved to be reasonably easy to use and reliably sensitive in various experimental settings in last twenty years [11].

Physiological Function Measurements. They are based on the symptoms that mental effort influences on physiological processes, such as oxygen consumption of brain [15], eye-blink and pupil dilation [16], p-wave of heart beat [17] and muscle tension [18].

Electrodermal Activity (EDA) was proved being sensitive to measure stress level [19] and often used in medical settings like nursery tasks [20]. Currently a wearable sensor, Affectiva Q sensor 2.0, can measure and record the EDA without interrupting of daily activities and causing discomfort [21].

Each category of methods measures a specific aspect of ML with one or two dozen of different parameters [4], so ML research in the last ten years developed multi-dimensional models to integrate these different methods, and these models usually base on expert rating, neural network and Multiple Resource Theory [22].

1.3 Mental Load Evaluation with Computer Systems

Currently, ML evaluation on human-computer interaction mainly focuses on VDTs. For instance, ML of arithmetic task on visual display terminal was firstly measured in 2006 [23]. Li et al. analyzed interaction ML in internet search and dual-task diagram in 2009, and combined factor analysis, back propagation neural network and self-organized neural network to establish a synthesis assessment model [24]. A 20-task navigation usability test and post-test NASA-TLX were applied to compare the ML of enhanced sound menu and visual menu on mobile terminal [25]. The electrodermal activity, electrocardiogram, photoplethysmo-graphy electroencephalogram were used as ML indicators for web browsing task [26]. According to these research cases, it is common to test typical tasks of a computer system with about 30 student participants under lab environment for Smartphone ML assessment.

The purpose of this experiment is to explore a comprehensive method for Smartphone ML, and compare different measurements to find out easy-to-use and sensitive indexes.

2 Method

2.1 Participants

This study was conducted with 33 college students (8 females, 25 males; average age 20.1, SD = 1.3), who received course credits for their participation. Participants separated into three groups: Novice, Average and Skilled users, according to a pre-test questionnaire on their knowledge on Smartphone usage [27], as shown in Table 1.

Table 1. Participants in three groups according to smartphone using proficiency

User group	Male	Female	In-total
Novice	2	3	5
Average	14	4	18
Skilled	9	1	10

The conditions of the pre-test was setting functions into levels easy, medium, hard and top by the degree of difficulty and frequency of usage. Then participants were

asked to fill in a questionnaire containing the questions on their experience in number of years of smartphone using and the different functions they know, in order to determine in which user group they belonged.

2.2 Research Design

The usage of Smartphone is a process of cognitive action, which mainly depends on user's perception and thinking abilities [5], so the ML supposed to including four main dimensions: (1) primitive performance measurement from usability test, (2) secondary tasks performance as environmental interference, (3) subjective ML scoring and (4) stress level from Electrodermal Activity (EDA).

The method from Donnell was applied on selecting secondary tasks and evaluated them with five indexes (sensitivity, diagnosis, interference, demands of manipulation and acceptance of operator) [28], see Table 2. The difficulty gradient between four main tasks was also checked in this pilot test. After analyzing validity and reliability of the pilot test, the experiment design was modified.

Table 2. Secondary task comparison in pilot test

Secondary tasks	Sensitive	Diagnosis	Interference	Demands of manipulation	Acceptance of operator
Beat rhythm	–	+	○	–	○
Time estimate	○	–	–	○	○
Words memory	○	–	–	○	○
Mental calculation	+	–	–	–	○
Random number memory	–	+	○	–	+

Note: “–” means unsuitable; “○” means neutral; “+” means suitable.

Error record of the observing researcher indicated the standard usage and alternative paths as criteria to record the usability problems.

2.3 Measurements

According to the Compound User Model [3], participants separated in novice, average and skilled user groups based on their experience. The novice users have less than a half year experience; the average users have around one and half year experience and know the basic functions of the Smartphone OS; the skilled users should have at least three years of experience and have knowledge on advance functions in Smartphone OS.

The ML of Smartphone was evaluated using the following categories of measurements [29]:

Performance of Use. Users make errors during using which may indicate that interface design challenges the cognitive and action capabilities of users. Besides, these errors also prolong the task duration. Thus, *the number of usability error* and the *task duration* are parameters of primitive task performance.

Secondary Tasks. The secondary task was *Random Number Memory* (RNM), which is asking the participants remember a set of random digits in less than one minute and recall them after task. Since chosen as secondary task, *Random Number Memory* (RNM) indicates the capability of short-term memory, which represents mental resource occupation of human brain. The fewer the memory of numbers after operation, it means the larger mental resource occupation of the just finished task. The less the digits remembered, the larger the mental load is.

Subjective Rating. NASA-TLX is a widely applied questionnaire to indicate general mental load with *mental demand*, *physical demand*, *temporal demand*, *performance*, *effort* and *frustration* indexes. It measures the ML perceived by operators. A higher score means a higher perceived subjective ML.

Physiological Function. *Electrodermal Activity* (EDA) closely relates to stress in mind, so EDA (in μ Siemens) represents the degree of nervous excitement and alertness levels. Thus, it could imply the degree of attention. Higher EDA means more concentrated mind status.

The basic assumption is that when the task difficulty increases, the user's mental load increases, causing an increase in the number of usability errors, a longer task duration, decline in short-term memory, increased EDA values, and higher score on mental demand, temporal demand, effort and frustration of NASA-TLX.

Moreover, users with different usage proficiency will show a different mental load distribution in four main tasks, that is: The novice users could have a higher ML than other user groups in all main tasks; the average users might have higher ML in and above medium tasks; skilled users may only experience high ML with high and top main tasks.

2.4 Tasks

In this study, the experiment conducted with fixed posture under a quiet indoor environment for easier operation and better experiment control [4]. At first, the participants attended a pre-test interview about their experience on Smartphone, mental and physical status, personality type, environmental distractions (e.g. ambient noise), and the inform consent alike. Then, they needed to wear Affectiva Q sensor 2.0 on his/her distal forearm, and relaxed in five minutes to get stable physiological signal (EDA at 32 Hz) as their baseline. Before the test started, the participants asked to remember seven random digits, and the number of correct answers was recorded as a benchmark of RNM. At last, they received an introduction on how to fill NASA-TLX Questionnaire and learn the basic configuration of the test device (Samsung Galaxy S III with Android 4.1 OS).

The experiment contained 17 fundamental functions of smartphone OS, which were divided into four main tasks from low, medium, high to top difficulty levels. Each main

task included 3–5 sub-tasks, e.g. “save the missed call and name it as XX”. Thus, the participants could finish each task chain in a similar duration, if they have no usability problem in their operation. Moreover, the design of task chains ensured that these features have no significant overlap in the operation path. The various interface elements of the smartphone operating system (e.g. functions, meta-interactions, icons, controls, interface structure) distributed relative evenly across each chain.

Each participant finished all main tasks in random sequence on the test device. When the participants were conducting the tasks, the researchers recorded their number of usability errors and task duration of each main task. Each main task was followed by a new set of RNM task. When participants completed the RNM task, they evaluated their subjective mental load of this main task using NASA-TLX questionnaire. There was a three-minute break between each main task. The total experiment lasted between 20 and 30 min.

3 Results of Experiment

3.1 User Performance

Task durations across main tasks with different difficulty level were compared and it was found out that task duration was increased at level of hard (shown in Table 3). The SD values of time duration were increasing on primitive performance across main tasks.

Table 3. Task duration in different task level

Task duration	Low	Medium	High	Top
Mean	156.37	233.37	256.00	322.11
SD	56.26	92.52	101.78	177.99
SE	10.83	17.81	19.59	34.26

The differences in duration between low and other main tasks were all significant ($p < 0.01$); the differences between top and all other main task were significant too ($p < 0.05$, paired samples T-test shown in Table 4).

Table 4. Task duration paired sample T-test in four chains

	Low	Medium	High	Top
Low	–	77.00**	99.63**	165.74**
Medium		–	22.63	88.74*
High			–	66.11*

Figure 1 shows that task duration across different users increased, when the main tasks became more difficult. The duration of the novices grew fastest. However, there

are only significant difference between novice and skilled users on top main task (mean difference 255.35, at $p < 0.05$ in one-way ANOVA).

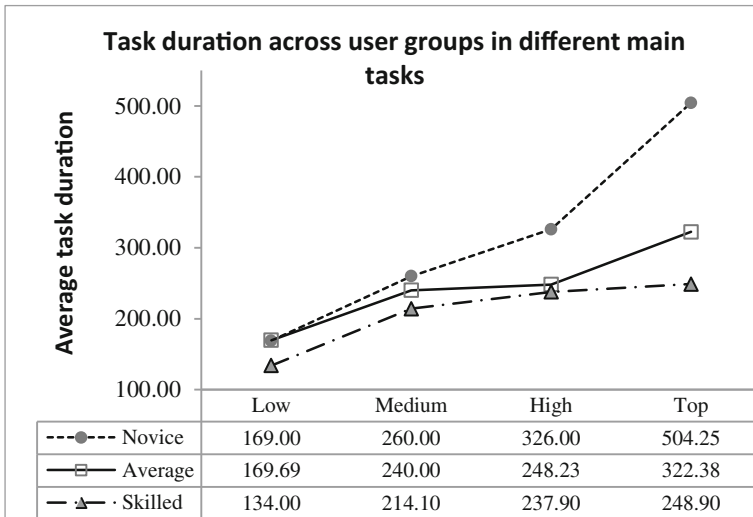


Fig. 1. Task durations of three users group across four chains.

As shown in Fig. 2, all users made more errors with the increasing difficulty of the main tasks ($p < 0.01$ in two-way t-test). The number of usability errors between novice and skilled users was significantly different, and the value between average and skilled users alike ($p < 0.05$ in paired t-test). For the skilled users, their error counts highly related to their task duration (Pearson $r_{sut} = 0.99$). In general, the number of usability errors and task duration had medium correlation (Pearson $r_{ut} = 0.69$).

3.2 Secondary Task

The pre-test and post-test numbers of RNM in different user groups were compared across main task (see Table 5 Memory ability pre and post each main task). The memory of novice and average users increased in lower level tasks, while decreased slightly after main tasks that were more difficult. The memory of skilled users just decreased evenly after the main tasks.

With the increasing difficulty, the RNM value decreased slightly (see Table 6), as subjective ML score also shown below.

3.3 Subjective Mental Load

The scores of skilled users were significantly higher than the novice and average users as shown in Table 7 ($p < 0.01$). The values of the *physical demand* index were significantly lower than other indexes, while the *performance* values were significantly the highest ($p < 0.05$ in t-test).

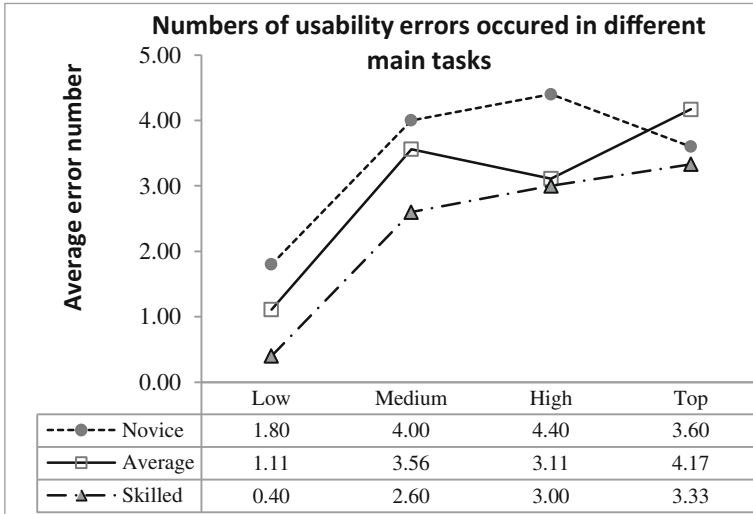


Fig. 2. The number of usability errors occurred in four main tasks.

Table 5. Memory ability pre and post each main task

User groups	Pre-task	Low	Medium	High	Top
Novice	4.80	5.20	4.60	3.80	4.40
Average	4.67	5.17	3.78	3.67	4.22
Skilled	5.40	4.10	4.20	4.20	4.30

Table 6. The proportion of users who has memory decline in three user groups

The proportion of users	Low	Medium	High	Top
Novice (%)	20.00	60.00	60.00	40.00
Average (%)	31.58	57.89	47.37	42.11
Skilled (%)	44.44	56.56	56.56	4.44

Table 7. NASA-TLX score of three user groups

	Mental demand	Physical demand	Effort	Frustration	Temporal demand	Performance
Novice	3.05	2.33	3.28	2.98	3.23	3.98
Average	3.50	2.72	3.40	2.90	3.61	4.32
Skilled	4.11	3.38	4.53	3.77	4.90	5.33
Average	3.56	2.81*	3.73	3.21	3.91	4.54*

As shown in Fig. 3, though the *physical demand* was low in Smartphone tasks, it still highly correlated to *mental demand* (Pearson $r_{\text{mndpd}} = 0.90$).

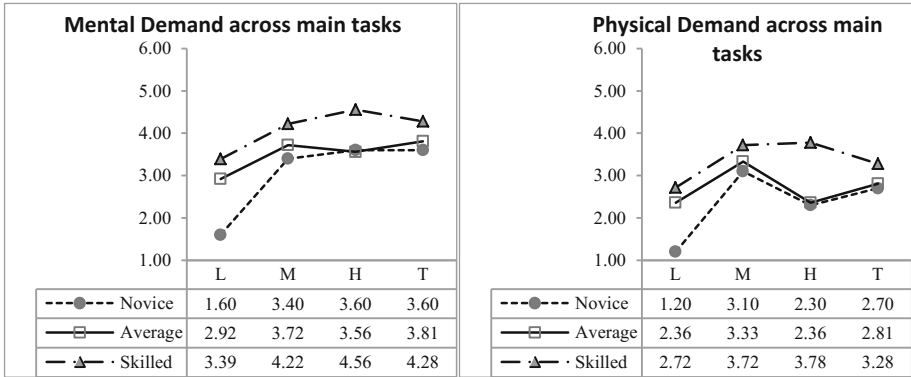


Fig. 3. The mental and physical demand of three user groups.

On *mental demand*, *frustration*, *effort* and *physical demand*, the skilled user’s subjective mental load increased slightly, when the main tasks became more difficult, as shown in Figs. 3 and 4. Moreover, the skilled user’s values on all these indexes highly related to the number of usability errors (Pearson $r_{Smd} = 0.95$, Pearson $r_{Sf} = 0.95$, Pearson $r_{Se} = 0.85$, Pearson $r_{Spm} = 0.79$).

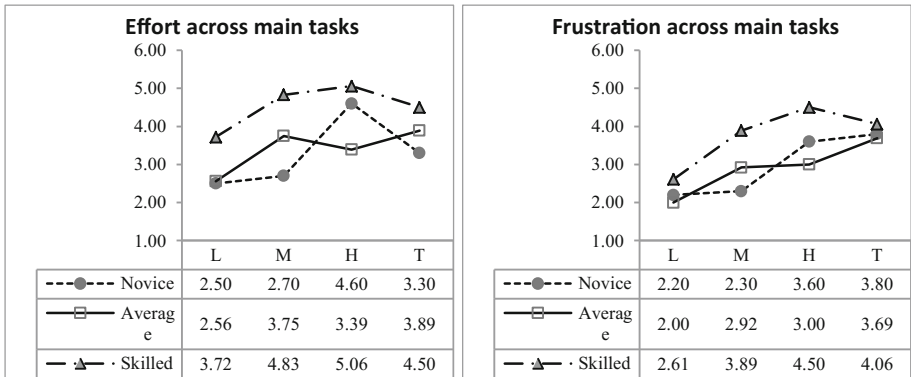


Fig. 4. The effort and frustration demand of three user groups.

The subjective mental load of both novice and skilled users fluctuated slightly across different main tasks. Similar to the skilled users, *mental demand* of average and novice users also highly related to their number of usability errors (Pearson $r_{Amd} = 0.99$, Pearson $r_{Nmd} = 0.95$). The novice user’s *physical demand* and their number of usability errors was also highly correlated (Pearson $r_{Npd} = 0.97$).

In general, the *effort* and *frustration* were relative highly correlated among users, as shown in Fig. 4 (Pearson $r_{ef} = 0.87$). However, only for the skilled and average users, their *effort* and *frustration* had high correlation to their number of usability errors (Pearson $r_{Ae} = 0.96$, Pearson $r_{Af} = 0.99$).

Like *mental demand*, *temporal demand* had also high correlation with the number of usability errors across different users (Pearson $r_{Std} = 0.94$, Pearson $r_{Atid} = 0.82$, Pearson $r_{Ntd} = 0.88$). However, the users rated their *performance* relatively low for simpler task chains, which was negatively related to their number of usability errors (Pearson $r_p = -0.81$). Detailed data is shown in Fig. 5.

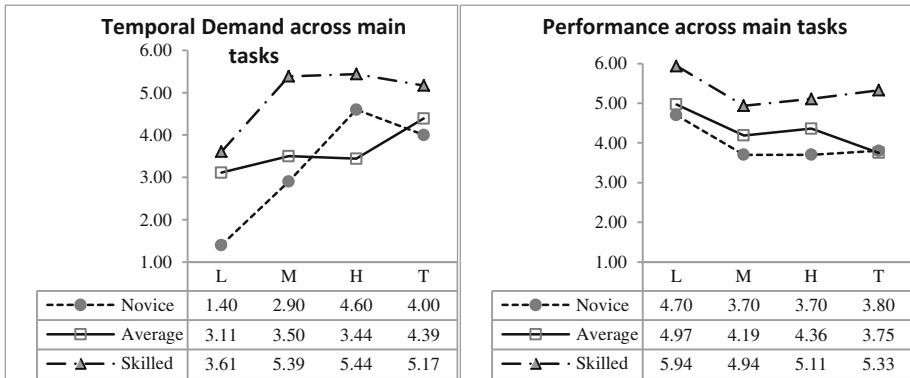


Fig. 5. The temporal demand and performance of three user groups.

3.4 Electrodermal Activity Values (EDA)

The value of the user's EDA varied from 0.04 to 24.91 μs , so the average EDA in different main tasks did not show significant difference across three user groups. Therefore, the minimal and maximal points of EDA were picked up in each main task. Although individual differences are large, but we can see a slow rise on EDA after difficulty increased (Fig. 6). Only at the minimal EDA level, the novice and skilled user show significant differences ($p < 0.05$ in t-test). Besides, there were also significant differences in low-high comparison and low-top comparison ($p < 0.05$ in t-test). The minimal EDA values had weak correlation (Pearson $r_{eda} = 0.30$) to the number of usability errors.

4 Discussion

Comparison of Secondary Task Methods. This study rated five secondary tasks in a pilot study, including beat rhythm, time estimate, words memory, mental calculation and random number memory, according to sensitivity, diagnosis, interference, demands of manipulation and acceptance of operators in a pilot test. The RNM had the highest acceptance value of operator. Therefore, it was chosen in order to control the mental load of the secondary task itself.

User Performance. In difficult main tasks, the novice users reduced their operating speed in order to avoid the errors. In the top task chain, because the novices consumed much more time, carefully learning concepts of advance functions at first, their task

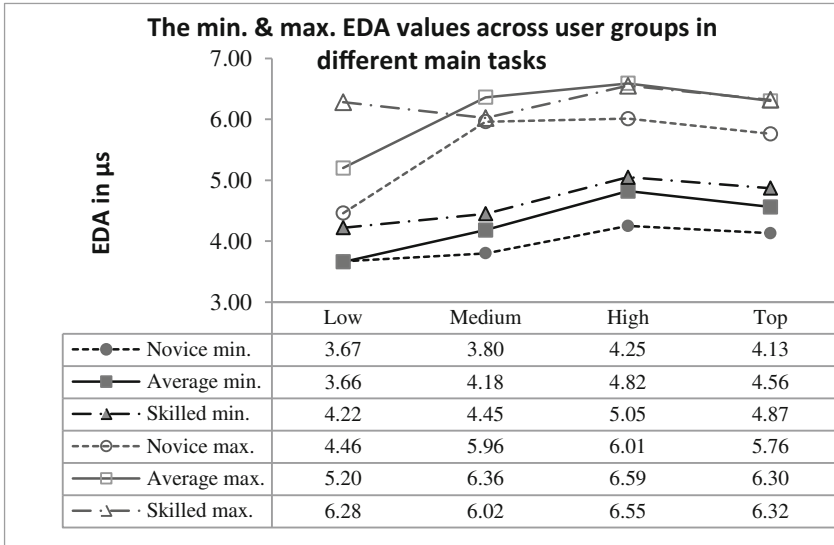


Fig. 6. The min. & max. EDA values of three user groups in four main tasks

duration increased very fast whereas the number of usability errors dropped. Moreover, the average user made more errors in medium level than high level, but not consumed much more time. This indicates that they feel confident adopting the “trial and error” strategy in familiar operation and could find the correct path fast. The main cause of increased task duration was the usability errors in operation due to their correlation coefficient; especially for skilled user they showed a near linear correlation. When the main tasks become more difficult, the performance differences across the users are larger, as indicated by the standard deviation increase on task duration.

Therefore, task duration could better reflect the ML of learning for novice users, while the number of usability errors is suitable to evaluate ML in familiar tasks for average and skilled user. Moreover, the number of usability errors also shows high correlation to subjective mental load.

Secondary Task. The memory of novice users and average users increased in lower level tasks, which may due to a higher degree of brain excitability. Since these tasks had no time requirement, short-term memory was not influenced by temporal stress. The rising of difficulty level in tasks was the main cause of growing mental effort. The skilled user had more memory loss might due to their higher self-expectation on task accomplishment, which consists of the high mental load on performance.

Thus, RMN easily reflects user’s self-expectation of their performance.

Subjective Scale Evaluation. The mental demand, effort, frustration and temporal demand indexes in NASA-TLX were all highly related to the number of usability errors. Especially for mental demand index, they are linear correlated across user groups.

Since the participants conducted all main tasks in a quiet lab with fixed sitting posture, the real physical demand was low. However, the correlation of *physical demand* for a slight physical task like operating a Smartphone, indicate that the perceived *physical demand* is highly influenced by *mental demand*.

Mental demand, *frustration* and *temporal demand* of the skilled users were all nearly linear correlated to their number of usability errors. For the average users, *mental demand* and *effort* showed the highest correlations to their number of usability errors. For the novice users, their *mental demand* strongly related to their *physical demand*.

On *performance* values, they had high negative correlation to the number of usability errors. Thus, it could be explained that the participants had higher self-expectation on task accomplishment in lab environment than in real using context, especially the skilled users. This experienced the psychological pressure could trigger a higher mental load.

EDA Measurement. In the pilot test, the influence of gender on the EDA was obvious, that female participants had higher EDA level than male, and the fluctuation was more drastic.

Actually, EDA indirectly related with mental load via user's emotional fluctuation. For the person who said s/he is extrovert in per-test interview, the relation between EDA and workload is not significant. For the person who is introvert, when pressure generated from tension and unconfident mood in more difficult tasks, the EDA fluctuation would become evident, thus the average EDA values would rise. Therefore, the EDA is more like a qualitative measurement for Smartphone ML.

However, there was an only a weak correlation between minimal EDA values and the number of usability errors. The small sample size of novice users may cause the high fluctuation on EDA values. Further research is needed with more participants to find out a specific relationship between EDA and Smartphone mental load.

5 Overall Conclusion

Though the mental load assessment for Smartphone is more complex than conventional usability tests, it could offer richer information on the correlation between different factors that may influence the quality of use.

Furthermore, different measurements could suit to different proficiency of users. For instance, task duration is a better proxy understanding the mental load of novice users, while the number of usability errors is better to evaluate the mental load of the average and skilled user.

Due to the high correlation between subjective scale and the number of usability errors, further researches could focus on simplify this test using less measurements with similar validity.

Besides, RNM and EDA are easily influenced by psychological pressure, which are more related to self-expectation and personality. Further researches could explore their relationships to Smartphone ML under time pressure.

References

1. ISO (2001) ISO/IEC 9126-1:2001 Software engineering – Product quality – Part 1: Quality model. International Standard. International Organization for Standardization, Switzerland
2. ISO: ISO9241-11(1998) 1998 Ergonomic requirements for office work with visual display terminals (VDT's) – Part 11: Guidance on usability. International Standard, International Organization for Standardization, Switzerland
3. Li LS (1999) Action theory and cognitive psychology in industrial design: User models and user interfaces. Dissertation, Art University of Braunschweig, Braunschweig
4. Kantowitz BH (1987) Mental Workload. In: Hancock PA. (ed) *Advances in Psychology*, vol 47, pp 81–121
5. Liao JQ (1995) Mental workload and its measurement. *J Syst Eng* 10(3):119–123
6. Kang WY, Yuan XG, Liu ZQ, Liu W (2008) Synthetic Evaluation method of mental workload on visual display interface in airplane cockpit. *Space Med Med Eng* 21(2):103–107
7. Li L, Yuan M (2011) Influential factors analysis of drivers' mental workload with the use of vehicle navigation system. *J Saf Environ* 11(6):202–204
8. Cui K, Sun LY, Feng TW, Xing X (2008) New developments in measurement methodologies of mental workload. *Industr Eng J* 11(5):1–5
9. Cooper GE, Harper RP, Jr (1969) The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities. Report No NASA TN-D-5153. Technical Report, Ames Research Center, National Aeronautics and Space Administration. Moffett Field
10. Hart SG (2006) NASA-task load index (NASA-TLX); 20 years later. In: *Proceedings of the human factors and ergonomics society 50th annual meeting*, vol 50. Sage Publications, Los Angeles, pp 904–908
11. Reid GB, Nygren TE (1988) The subjective workload assessment technique: a scaling procedure for measuring mental workload. *Adv Psychol* 52:185–218 Elsevier Science Publishers, North Holland
12. Boles DB, Bursk JH, Phillips JB, Perdelwitz JR (2007) Predicting dual-task performance with the multiple resources questionnaire (MRQ). *Hum Factors* 49(1):32–45
13. Galy E, Cariou M, Mélan C (2011) What is the relationship between mental workload factors and cognitive load types? *Int J Psychophysiol* 83(3):269–275
14. Shingledecker CA, Crabtree MS, Simons JC et al (1980) Subsidiary Radio Communications Tasks for Workload Assessment in R&D Simulations I. Task Development and Workload Scaling. Technical Report, Systems Research Labs Inc, Dayton Ohio
15. Horst RL, Johnson R, Donchin E (1980) Event-related brain potentials and subjective probability in a learning task. *Mem Cognit* 8(5):476–488
16. Ahlstrom U, Friedman-Berg FJ (2006) Using eye movement activity as a correlate of cognitive workload. *Int J Industr Ergon* 36(7):623–636
17. Gunn CG, Wolf S, Block RT et al (1972) Psychophysiology of the cardiovascular system. In: Greenfield NS, Sternbach RA (eds) *Handbook of psychophysiology*. Holt, Rinehart & Winston, New York, pp 457–483
18. Suzuki S, Kumano H, Sakano Y (2003) Effects of effort and distress coping processes on psychophysiological and psychological stress responses. *Int J Psychophysiol* 47(2):117–128
19. Reinhardt T, Schmahl C, Wüst S, Bohus M (2012) Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the manheim multicomponent stress test (MMST). *Psychiatry Res* 198(1):106–111

20. Moya-Albiol L, Sanchis-Calatayud MV, Sariñana-González P, De Andrés-García S, Romero-Martínez Á, González-Bono E (2012) P03-425 - Electrodermal activity in response to a set of mental tasks in caregivers of persons with autism spectrum disorders. *Eur Psychiatry* 26(1):1595
21. Affectiva (2012) Liberate yourself from the lab: Q Sensor measures EDA in the wild. Affectiva QTM Solutions White Paper
22. Wang J, Fang WN, Li GY (2010) Mental workload evaluation method based on multi-resource theory model. *J. Beijing Jiaotong Univ.* 34(6):107–110
23. Peng XW, He QC, Ji T, Wang ZL, Yang L (2006) Mental workload for mental arithmetic on visual display terminal. *Chin J Industr Hyg Occup Dis* 24(12):726–729
24. Li JB, Xu BH (2009) synthetic assessment of cognitive load in human-machine interaction process. *Acta Psychologica Sinica* 41(1):35–43
25. Yu YH, Li ZJ (2011) Study of sonically enhanced menu interaction for mobile terminals. *Appl Res Comput* 28(10):3742–3745
26. Jimenez-Molina A, Retamal C, Lira H (2018) Using psychophysiological sensors to assess mental workload during web browsing. *Sensors* 18(2):458
27. Li M (2008) Comparison of Usability Evaluation Method Based-on Needs of Software Development. Master Thesis, Xi'an Jiaotong University, Xi'an
28. O'Donnell RD, Eggemeier FT (1986) Workload assessment methodology. In: Boff KR, Kaufman L, Thomas JP (eds) *Handbook of perception and human performance*, vol II. Wiley, New York, pp 42–43
29. Galy E, Cariou M, Mélan C (2012) What is the relationship between mental workload factors and cognitive load types? *Int J Psychophysiol* 83(3):269–275