Human Interaction and Emerging Technologies (IHIET-AI 2024), Vol. 120, 2024, 305–315
https://doi.org/10.54941/ahfe1004581

The Convergent Validity of Computer Operating Systems' Usability Evaluation by Popular Generative Artificial Intelligence (AI) Robots

Victor K. Y. Chan

Faculty of Business, Macao Polytechnic University, Macao, China

ABSTRACT

This article seeks to examine the convergent validity of (and thus the consistency between) computer operating systems' (OSs') usability evaluation by a number of popular generative artificial intelligence (AI) robots. Totally 18 popular OS versions were included in the study, they specifically being the various versions of the three leading OS families of Windows, macOS, and Linux. Usability was evaluated in eight major dimensions, namely, (1) effectiveness, (2) efficiency, (3) learnability, (4) memorability, (5) safety, (6) utility, (7) ergonomics, and (8) accessibility. Experimenting with a handful of generative AI robots, Microsoft's Copilot, Google's PaLM, and Meta's Llama managed to individually accord rating scores to the aforementioned eight dimensions. For each robot of this trio, the minimum, the maximum, the range, and the standard deviation of the rating scores for each of the eight dimensions were computed across the OS versions. The rating score difference for each of the eight dimensions between each pair of these robots was calculated for each OS version. The mean of the absolute value, the minimum, the maximum, the range, and the standard deviation of the differences for each dimension between each robot pair were calculated across the OS versions. A paired sample *t*-test was then applied to each dimension for the rating score difference between each robot pair over the versions. Finally, Cronbach's coefficient alpha (α) of the rating scores was computed for each dimension between all the three robots across the versions. These computational outcomes were to affirm whether each robot awarded discrimination in evaluating each dimension across the OS versions, whether each robot vis-à-vis any other robots erratically and/or systematically overrate or underrate any dimension over the OS versions, and whether there was high convergent validity of (and thus consistency between) all the three robots in evaluating each dimension across the OS versions. Among other ancillary results, it was found that the convergent validity of the three robots in evaluating all the eight dimensions was high, and thus such evaluation is trustworthy at least to an extent.

Keywords: Artificial intelligence, Robots, Usability, Computer operating system versions, Convergent validity

INTRODUCTION

The usability of computer operating systems (OSs) is an important factor that affects the user experience, productivity, and satisfaction of the users. Usability can be defined as "the extent to which a system, product or service

can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO, 2018). Traditionally, the usability of OSs has been evaluated by usability testing through manual methods such as questionnaire surveys, interviews, and focus groups (Maramba, Chatterjee, and Newman, 2019), which involve collecting data from the users or experts and analyzing them to identify the strengths and weaknesses of the OS design (Nielsen, 1994). However, these methods have some limitations, such as being time-consuming, costly, subjective, and dependent on the availability and representativeness of the participants.

Generative artificial intelligence (AI) is a branch of AI that aims to create new content or data based on data that the AI systems have been trained on. Generative AI creates new content in the form of images, text, audio, and more (Baidoo-Anu and Ansah, 2023; Gartner, 2023; World Economic Forum, 2023). Generative AI can serve as an alternative method to evaluate the usability of OSs by, for example, generating realistic scenarios, tasks, or user feedback that can simulate the actual use of the OSs by different types of users. In addition, generative AI robots can be deployed to survey the opinions about a large variety of OSs in a swift, inexpensive, and compendious manner. In particular, such surveys can be conducted readily on the Internet, which is populated by profuse opinions of OS users. Generative AI has some advantages over traditional methods of evaluating OSs' usability, such as being faster, cheaper, more objective, and more scalable.

In the last decade, the employment of generative AI robots to investigate phenomena concerning OSs has been broadly examined. In the realm of mobile device operating systems, for example, Amin et al. (2022) put forward a technique to cater for malware detection, which was by design a deep learning model making use of generative adversarial networks. It was responsible for detecting Android malware by means of famous two-player game theory for a rock-paper-scissor problem. The researchers used three state-of-the-art datasets and a large-scale dataset of opcodes extracted from the Android Package Kit bytecodes. The technique achieved an F1 score of 99% with a receiver operating characteristic of 99% on the bytecode dataset.

Another example is that Huang et al. (2022) proposed Android-SEM, which was an Android source code semantic enhancement model based on transfer learning. The proposed model was built upon the Transformer architecture to achieve a pre-training framework for generating code comments from malware source codes. The performance of the pre-training framework was optimized using a generative adversarial network. The proposed model relied on a novel regression model-based filter to retain high-quality comments and source codes for feature fusion pertinent to semantic enhancement. Creatively, and contrary to conventional methods, a quantum support vector machine (QSVM) was incorporated for classifying malicious Android codes by combining quantum machine learning and classical deep learning models. The results proved that Android-SEM achieved accuracy levels of 99.55% and 99.01% for malware detection and malware categorization, respectively.

Notwithstanding, the author is not cognizant of any existing literature focusing on the evaluation of OSs' usability, whether for mobile devices or computers in general, by means of generative AI robots. This is exactly

the gap that this article is to fill. In particular, this article seeks to examine the convergent validity of (and thus the consistency between) OSs' usability evaluation by a number of popular generative AI robots and covers the versions of the three leading OS families of Microsoft Windows, Apple macOS, and Linux. The 18 versions so included are Windows XP, Windows Vista, Windows 7, Windows 8, Windows 10, Windows 11, Mac OS X, OS X, macOS, Linux Mint, Manjaro Linux, Debian Linux, Ubuntu, Antergos/EndeavourOS, Solus, Fedora, elementary OS, and openSUSE and were shortlisted simply by referencing some official and authoritative websites of the three OS families, for example, Britannica, T. Editors of Encyclopaedia (2023) for the Windows family, Apple (2023) for the macOS family, and Anonymous (2023) for Linux.

METHODOLOGY

Data and Materials

The present study started off in November 2023 experimenting with four very popular generative AI robots, namely, Microsoft Copilot (Cambon et al., 2023), Assistant (Anonymous, 2024), Google PaLM (Anil et al., 2023), and Meta Llama (Oxford Analytica, 2023) as candidates for the evaluation of OSs' usability, the first one being bundled with the Microsoft Edge browser whereas the rest having been incorporated into the AI portal poe.com. Eight major dimensions to evaluate the perceived usability of any OS were identified as (1) Effectiveness (Chan, 2023; Raptis et al., 2013), (2) Efficiency (Chan, 2023; Raptis et al., 2013), (3) Learnability (Chan, 2023; Thillaieaswaran and Pasupathy, 2021), (4) Memorability (Bakiu and Guzman, 2017; Saket, Endert, Stasko, 2016), (5) Safety (Gurbuz and Tekinerdogan, 2018), (6) Utility (Okumuş et al., 2016), (7) Ergonomics (Peres et al., 2009), and (8) Accessibility (Bi et al., 2022; Chan, 2023), which were to be rated by the robots in this study. Effectiveness refers to the chance of users completing tasks successfully and correctly on an OS. Effectiveness of an OS can be measured by metrics such as users' task success rate and the number of errors made by users on the OS. Users' task success rate is high and users' number of errors is low for OSs with high effectiveness. Efficiency refers to the speed and accuracy with which users can complete their tasks using an OS. It is influenced by factors such as the speed of the system, responsiveness to user inputs, and the design of the user interface. Efficiency of an OS can be measured by metrics such as the average time for users to perform a certain number of specified tasks on the OS (Nielsen, 1994). Users complete tasks fast on OSs with high efficiency. Learnability refers to the ease with which users can learn how to use an OS. It is influenced by factors such as the availability of documentation, the simplicity of the user interface, and the consistency of the user interface. The learnability of an OS can be measured by metrics such as the time for novice users to reach a specified level of proficiency in using the OS or for them to be able to complete a certain task successfully on the OS (Nielsen, 1994). Novice users' time to be able to complete a certain task successfully on OSs with high learnability is short. Memorability is the ability of users to remember how to use an OS after a period of non-use. A

memorable OS should have a simple and memorable interface that minimizes the cognitive load and recall effort of users. Safety is the extent to which an OS protects users from errors, failures, threats, and harms. A safe OS should have a robust and reliable performance that prevents or recovers from system crashes, data loss, corruption, malware attacks, or unauthorized access. Utility is the range and quality of functionality and features that an OS provides to users. A high-utility OS should have a comprehensive and diverse set of applications and services that cater to different user needs and preferences, as well as support interoperability and compatibility with other devices and platforms. Ergonomics is the physical comfort and ease of use that an OS provides to users. An ergonomic OS should have a suitable design and layout that adapts to different user characteristics, such as the age, gender, culture, language, vision, hearing, and motor skills, as well as different environmental conditions, such as lighting, noise, temperature, and humidity. Accessibility is the degree to which an OS can be used by people with disabilities or special needs. An accessible OS should have a flexible and adaptable interface that supports various input and output methods, such as keyboard, mouse, touch screen, voice recognition, speech synthesis, braille display, or screen reader.

Then, the following instruction for rating scores was submitted to each of the four robots:

> "For each of the eight dimensions (1) Effectiveness, (2) Efficiency, (3) Learnability, (4) Memorability, (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility, please give a rating score to each of the major computer operating system versions Windows XP, Windows Vista, Windows 7, Windows 8, Windows 10, Windows 11, Mac OS X, OS X, macOS, Linux Mint, Manjaro Linux, Debian Linux, Ubuntu, Antergos/EndeavourOS, Solus, Fedora, elementary OS, openSUSE based on a scale of 1 to 10 (1 being the worst and 10 the best). Please derive your scores from global users' textual comments on these eight dimensions of these versions as appear all around the web. It would be nice if you put your scores in a table form."

With some hiccups, Copilot replied with rating scores for all the eight dimensions and the 18 OS versions enumerated in the above instruction but the rating scores for the dimensions (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility corresponding to the OS versions elementary OS and openSUSE were missing, on which the robot claimed that it could not find enough user comments to derive reliable scores. PaLM yielded the rating scores for all the eight dimensions and all the 18 OS versions. As for Llama, rating scores for all the eight dimensions but only 15 OS versions above were rendered with Fedora, elementary OS, and openSUSE skipped. On the contrary, Assistant predicated inability to provide any rating scores. In summary, it transpired that only the rating scores of Copilot, PaLM, and Llama were amenable to further analysis. Please note that the instruction above expressly underlined "…derive your scores from global users' textual comments on these eight dimensions of these versions as appear all around the web." In other words, each robot supposedly devised its rating scores from global users' textual

comments appearing all around the worldwide web instead of simply ditto-ing any extant rating scores of a similar nature assigned earlier by somebody or some robots else.

Analysis

For each of the three robots Copilot, PaLM, and Llama, the minimum, the maximum, the range, and the standard deviation of the rating scores for each of the eight dimensions were calculated across all the 18 (for the dimensions (1) Effectiveness, (2) Efficiency, (3) Learnability, and (4) Memorability in the case of Copilot and all the eight dimensions in the case of PaLM), 16 (for the dimensions (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility in the case of Copilot), and 15 (in the case of Llama) OS versions. If there is a substantial range and standard deviation for a particular dimension, it is affirmed that the corresponding robot accords discrimination in rating the dimension across the OS versions.

Subsequently, the rating score difference for each of the eight dimensions between any pair of robots was computed for each of the OS versions to which both robots in the pair accorded rating scores. The mean of the absolute values, the minimum, the maximum, the range, and the standard deviation of the differences for each dimension between each pair of robots were computed across all those OS versions with rating scores from both robots in the pair. If the mean of the absolute values, the range, and the standard deviation for a particular dimension are sufficiently small, it is signified that the robots in the pair neither overrate nor underrate erratically with respect to each other the dimension across the OS versions. A paired sample *t*-test was then applied to each dimension for the rating score differences between each robot pair over those OS versions with rating scores from both robots in the pair. If the *t*-test is significant for a particular dimension and the corresponding mean difference is positive (negative), it is verified that the first robot in the pair systematically overrates (underrates) the dimension with respect to the second robot.

Finally, for more statistically rigorous measurement of the consistency between all the three robots' evaluation, Cronbach's coefficient alpha (α) (DeVellis, 2005) of the rating scores was computed for each of the eight dimensions between all the three robots across all the OS versions to which all the three robots awarded rating scores. If Cronbach's coefficient alpha is high, for instance, over 0.5 or 0.6 (Ling et al., 2021; Nunnally, 1967) for a particular dimension, it is indicated that there is consistency between all the three robots in rating the dimension across all those OS versions with rating scores from all the three robots. Stated differently, the corresponding convergent validity of all the three robots in rating the dimension across such OS versions is high.

RESULTS

Table 1 lists the minimum, the maximum, the range, and the standard deviation of the rating scores as rated by each of the three robots for each of the eight dimensions across all the 18 (for the dimensions (1) Effectiveness, (2) Efficiency, (3) Learnability, and (4) Memorability in the case of Copilot

and all the eight dimensions in the case of PaLM), 16 (for the dimensions (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility in the case of Copilot), and 15 (in the case of Llama) OS versions. Whereas all the three robots rated with appreciable discrimination, Copilot and Llama did more so than PaLM, especially, in the four dimensions (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility in the case of Copilot as manifested by the disparity between the ranges and the standard deviations of these four dimensions' scores as rated by Copilot and Llama and those of other dimensions as also rated by Copilot and between the ranges and the standard deviations of most dimensions' scores as rated by Copilot and those as rated by PaLM.

Table 1. The minimum, the maximum, the range, and the standard deviation of the rating scores as rated by each of the three robots for each of the eight dimensions across all the 18 (for the dimensions (1) Effectiveness, (2) Efficiency, (3) Learnability, and (4) Memorability in the case of Copilot and all the eight dimensions in the case of PaLM), 16 (for the dimensions (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility in the case of Copilot), and 15 (in the case of Llama) OS versions.

Robot (sample size n)	Minimum/ maximum/ range/ standard deviation	Effectiveness	Efficiency	Learnability	Memorability	Safety	Utility	Ergonomics	Accessibility
Copilot ($n = 18$ or 16)	Minimum	5	4	4	4	2	3	4	3
	Maximum	9	8	7	7	9	9	9	9
	Range	4	4	3	3	7	6	5	6
	Standard deviation	1.0556	1.0556	0.7859	0.8324	1.8875	1.6820	1.4009	1.6279
PaLM ($n = 18$)	Minimum	6	6	5	5	6	6	6	6
	Maximum	9	9	9	9	9	9	9	9
	Range	3	3	4	4	3	3	3	3
	Standard deviation	0.8498	0.8498	0.9164	0.9164	0.7670	0.7670	0.7670	0.7670
Llama ($n = 15$)	Minimum	6	5	4	3	7	6	5	4
	Maximum	9	9	8	7	10	10	9	8
	Range	3	4	4	4	3	4	4	4
	Standard deviation	0.9759	1.1952	1.2228	1.2228	0.8837	1.2228	1.1952	1.2228

Table 2 enumerates the mean of the absolute values, the minimum, the maximum, the range, and the standard deviation of the rating score differences for each of the eight dimensions between each pair of robots across all those OS versions with rating scores from both robots in the pair. In comparison to PaLM, Copilot tended to have overrated or underrated less erratically the dimensions (1) Effectiveness and (2) Efficiency in view of the corresponding means of the absolute values, the corresponding ranges, and the corresponding standard deviations of the differences being less than those for all the other six dimensions in respect of this robot pair and than those for all dimensions in respect of other robot pairs.

Table 2. The mean of the absolute values, the minimum, the maximum, the range, and the standard deviation of the rating score differences for each of the eight dimensions between each pair of robots across all those OS versions with rating scores from both robots in the pair.

Robot pair (sample size *n*)	Mean of the absolute values/ minimum/ maximum/ range/ standard deviation of the differences	Effectiveness	Efficiency	Learnability	Memorability	Safety	Utility	Ergonomics	Accessibility
Copilot – PaLM ($n = 18$[a] or 16[b])	Mean of the absolute values	0.4444	0.6667	1.7778	1.7222	1.1875	0.8125	0.8125	1.25
	Minimum	−1	−2	−3	−3	−5	−3	−2	−4
	Maximum	1	0	−1	0	1	2	2	1
	Range	2	2	2	3	6	5	4	5
	Standard deviation	0.5941	0.5941	0.7321	0.8948	1.4361	1.2633	1.0782	1.3663
Copilot – Llama ($n = 15$)	Mean of the absolute values	1.0667	1	1.0667	1.2	1.6667	1.0667	0.8	1.0667
	Minimum	−2	−2	−2	−1	−6	−3	−1	−2
	Maximum	3	3	2	3	0	1	2	3
	Range	5	5	4	4	6	4	3	5
	Standard deviation	1.3345	1.3870	1.2536	1.2228	1.7593	1.2649	1.0328	1.2910
PaLM – Llama ($n = 15$)	Mean of the absolute values	0.9333	1.0667	1.7333	2.6	1.0667	1	1.1333	1.8
	Minimum	−2	−2	−1	0	−3	−3	−2	−1
	Maximum	2	3	4	5	1	2	3	4
	Range	4	5	5	5	4	5	5	5
	Standard deviation	1.2459	1.3558	1.3522	1.3522	1.1751	1.2910	1.3522	1.2910

[a] For the dimensions (1) Effectiveness, (2) Efficiency, (3) Learnability, and (4) Memorability
[b] For the dimensions (5) Safety, (6) Utility, (7) Ergonomics, and (8) Accessibility

Table 3 details the paired sample *t*-tests of the rating score differences for each of the eight dimensions between each pair of robots over all those OS versions with rating scores from both robots in the pair. Relative to PaLM, Copilot inclined to systematically overrate the dimension (1) Effectiveness (at the 5% significance level or $p < 0.05$) and underrate the five dimensions (2) Efficiency ($p < 0.01$), (3) Learnability ($p < 0.01$), (4) Memorability ($p < 0.01$), (5) Safety ($p < 0.05$), and (8) Accessibility ($p < 0.05$) whereas Llama tended to systematically underrate the three dimensions (3) Learnability ($p < 0.01$), (4) Memorability ($p < 0.01$), and (8) Accessibility ($p < 0.01$) and overrate (5) Safety ($p < 0.05$). With respect to Llama, Copilot also inclined to systematically overrate (4) Memorability ($p < 0.01$) and underrate (5) Safety ($p < 0.01$) and (6) Utility ($p < 0.05$). Otherwise, vis-à-vis each other, the three robots neither overrated nor underrated systematically any other dimensions.

Table 3. The paired sample *t*-test of the rating score differences for each of the eight dimensions between each pair of robots over all those OS versions with rating scores from both robots in the pair.

Differences (sample size *n*)		Dimension	Mean difference / [95% confidence interval]	*t* (*p*-value) / degrees of freedom
Copilot – PaLM	($n = 18$)	Effectiveness	.333 / [.038,.629]	2.380 (.029*) / 17
		Efficiency	−.667 / [−.962, −.371]	−4.761 (.000**) / 17
		Learnability	−1.778 / [−2.142, −1.414]	−10.303 (.000**) / 17
		Memorability	−1.722 / [−2.167, −1.277]	−8.166 (.000**) / 17
	($n = 16$)	Safety	−.937 / [−1.703, −.172]	−2.611 (.020*) / 15
		Utility	−.437 / [−1.111,.236]	−1.385 (.186) / 15
		Ergonomics	−.312 / [−.887,.262]	−1.159 (.264) / 15
		Accessibility	−1.000 / [−1.728, −.272]	−2.928 (.010*) / 15
Copilot – Llama ($n = 15$)		Effectiveness	.267 / [−.472, 1.006]	.774 (.452) / 14
		Efficiency	−.067 / [−.835,.701]	−.186 (.855) / 14
		Learnability	.000 / [−.694,.694]	.000 (1.000) / 14
		Memorability	1.067 / [.390, 1.744]	3.378 (.005**) / 14
		Safety	−1.667 / [−2.641, −.692]	−3.669 (.003**) / 14
		Utility	−.800 / [−1.500, −.100]	−2.449 (.028*) / 14
		Ergonomics	.267 / [−.305,.839]	1.000 (.334) / 14
		Accessibility	.667 / [−.048, 1.382]	2.000 (.065) / 14
PaLM – Llama ($n = 15$)		Effectiveness	−.133 / [−.823,.557]	−.414 (.685) / 14
		Efficiency	.533 / [−.217, 1.284]	1.524 (.150) / 14
		Learnability	1.600 / [.851, 2.349]	4.583 (.000**) / 14
		Memorability	2.600 / [1.851, 3.349]	7.447 (.000**) / 14
		Safety	−.667 / [−1.317, −.016]	−2.197 (.045*) / 14
		Utility	−.333 / [−1.048,.382]	−1.000 (.334) / 14
		Ergonomics	.600 / [−.149, 1.349]	1.718 (.108) / 14
		Accessibility	1.667 / [.952, 2.382]	5.000 (.000**) / 14

* $p < 0.05$; ** $p < 0.01$

Table 4 depicts Cronbach's coefficient alpha of the rating scores for each of the eight dimensions between all the three robots over those 15 OS versions with rating scores from all the three robots. All the eight dimensions yielded values of Cronbach's coefficient alpha high enough (Ling et al., 2021; Nunnally, 1967) to infer consistency between the three robots in evaluating the dimensions. In summary, the convergent validity of the three robots was "more than" acceptable for all the eight dimensions, and thus the three robots may be rather, if not absolutely, trustworthy in evaluating all the eight dimensions of the OS versions' usability.

Table 4. Cronbach's coefficient alpha of the rating scores for each of the eight dimensions between all the three robots over those 15 OS versions with rating scores from all the three robots.

Sample size *n*	Effectiveness	Efficiency	Learnability	Memorability	Safety	Utility	Ergonomics	Accessibility
15	.672	.689	.657	.638	.626	.764	.754	.725

CONCLUSION AND DISCUSSION

With a view to assessing the convergent validity of (and thus the consistency between) generative AI robots in evaluating OS versions' usability, it was uncovered that the convergent validity of the three robots Copilot, PaLM, and Llama was "more than" acceptable for all the eight usability dimensions.

Having said that, there are still quite some factors precipitating inconsistency, however small, between robots in the evaluation of OS versions' usability (or, in fact, anything under the sun), for example, the subjectivity inherent in the textual user comments on which the robots were trained and the presumably disparate samples of textual user comments accessible to different robots and on which different robots were trained (Chan, 2023).

This study itself is not without its critics. First, this study experimented with only three generative AI robots Copilot, PaLM, and Llama, which might not be able to epitomize the profuse robots globally. Second, these three robots were trained on data samples up to a few years back, so even the rating scores from them today may not be indicative of the latest OS versions' usability. Therefore, future researches of purposes comparable to this study may augment the set of generative AI robots, especially, those having been trained on the latest data samples.

In spite of these limitations, generative AI robots are emerging as a promising and transformative method to insightfully comprehend global users' textual comments at scale so as to rate the diverse usability dimensions of each OS. Such rating is far faster, less costly, more objective, and more inclusive in the coverage of opinions from users of various geographic locales worldwide than virtually any manual methods.

REFERENCES

Amin, Muhammad, Shah, Babar, Sharif, Aizaz, Ali, Tamleek, Kim, Ki-Il, and Anwar, Sajid. (2022). Android Malware Detection through Generative Adversarial Networks, Transactions on Emerging Telecommunications Technologies Volume 33 No. 2. Website: https://doi.org/10.1002/ett.3675

Anil, Rohan, Dai, Andrew M., Firat, Orhan, Johnson, Melvin, Lepikhin, Dmitry, Passos, Alexandre, et al. (2023). PaLM 2 Technical Report. Website: https://doi.org/10.48550/arXiv.2305.10403

Anonymous. (2023). What Is Linux? Website: https://www.linux.com/what-is-linux/

Anonymous. (2024). Poe. Website: https://poe.com/chats

Apple. (2023). Find out which macOS your Mac is using. Website: https://support.apple.com/en-hk/HT201260

Baidoo-Anu, D. and Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning, Journal of AI Volume 7 No. 1. Website: https://dergipark.org.tr/en/pub/jai/issue/77844/1337500

Bakiu, E. and Guzman, E. (2017). "Which Feature is Unusable? Detecting Usability and User Experience Issues from User Reviews", proceedings of 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW), Lisbon, Portugal. Website: 10.1109/REW.2017.76

Bi, Tingting, Xia, Xin, Lo, David, Grundy, John, Zimmermann, Thomas, and Ford, Denae. (2022). Accessibility in Software Practice: A Practitioner's Perspective. ACM Transactions on Software Engineering and Methodology Volume 31 No. 4. Website: https://doi.org/10.1145/3503508

Britannica, T. Editors of Encyclopaedia. (2023). Microsoft Windows, Encyclopedia Britannica. Website: https://www.britannica.com/technology/Microsoft-Windows

Cambon, Alexia, Hecht, Brent, Edelman, Ben, Ngwe, Donald, Jaffe, Sonia, Heger, Amy, Vorvoreanu, Mihaela, Peng, Sida, Hofman, Jake, Farach, Alex, Bermejo-Cano, Margarita, Knudsen, Eric, Bono, James, Sanghavi, Hardik, Spatharioti, Sofia, Rothschild, David, Goldstein, Daniel G., Kalliamvakou, Eirini, Cihon, Peter, Demirer, Mert, Schwarz, Michael, and Teevan, Jaime. (2023). Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. Website: https://www.microsoft.com/en-us/research/uploads/prod/2023/12/AI-and-Productivity-Report-First-Edition.pdf

Chan, Victor K. Y. (2023). The Consistency between Popular Generative Artificial Intelligence (AI) Robots in Evaluating the User Experience of Mobile Device Operating Systems, Artificial Intelligence and Social Computing Volume 113.

DeVellis, Robert. F. (2005). "Inter-rater reliability", in: Encyclopedia of social measurement, Kempf-Leonard, K. pp. 317–322.

Gartner. (2023). Gartner Experts Answer the Top Generative AI Questions for Your Enterprise: Generative AI isn't just a Technology or a Business Case—it is a Key Part of a Society in Which People and Machines Work Together. Website: https://www.gartner.com/en/topics/generative-ai

Gurbuz, H. G. and Tekinerdogan, B. (2018). Model-Based Testing for Software Safety: a Systematic Mapping Study, Software Quality Journal Volume 26. Website: https://doi.org/10.1007/s11219-017-9386-2

Huang, Yizhao, Li, Xingwei, Qiao, Meng, Tang, Ke, Zhang, Chunyan, Gui, Hairen, Wang, Panjie, and Liu, Fudong. (2022) Android-SEM: Generative Adversarial Network for Android Malware Semantic Enhancement Model Based on Transfer Learning, Electronics Volume 11 No. 5. Website: https://doi.org/10.3390/electronics11050672

ISO. (2018). ISO 9241-11:2018 Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts. Website: https://www.iso.org/obp/ui/en/#iso:std:iso:9241:-11:ed-2:v1:en

Ling, Hsiao-Chi, Chen, Hong-Ren, Ho, Kevin K. W., Hsiao, Kuo-Lun (2021). Exploring the Factors Affecting Customers' Intention to Purchase a Smart Speaker, Journal of Retailing and Consumer Services Volume 59.

Maramba, I., Chatterjee A., and Newman, C. (2019). Methods of Usability Testing in the Development of eHealth Applications: A Scoping Review, International Journal of Medical Informatics Volume 126. Website: https://doi.org/10.1016/j.ijmedinf.2019.03.018.

Nielsen, J. (1994). Usability Engineering. San Francisco: Morgan Kaufmann Publishers.

Nunnally, J. C. (1967). Psychometric Theory, McGraw-Hill.

Okumuş, S., Lewis, L., Wiebe, E., and Hollebrands, K. (2016). Utility and Usability as Factors Influencing Teacher Decisions about Software Integration, Educational Technology Research and Development Volume 64. Website: https://doi.org/10.1007/s11423-016-9455-4

Oxford Analytica. (2023). Meta LLaMa Leak Raises Risk of AI-Linked Harms. Expert Briefings. Website: https://doi.org/10.1108/OXAN-ES276597

Peres, S. Camille, Nguyen, Vickie, Kortum, Philip T., Akladios, Magdy, Wood, S. Bart, and Muddimer, Andrew. (2009). "Software Ergonomics: Relating Subjective and Objective Measures", proceedings of CHI '09 Extended Abstracts on Human Factors in Computing Systems. Website: https://doi.org/10.1145/1520340.1520599

Raptis, Dimitrios, Tselios, Nikolaos, Kjeldskov, Jesper, and Skov, and Mikael. B. (2013). "Does Size Matter? Investigating the Impact of Mobile Phone Screen Size on Users' Perceived Usability, Effectiveness and Efficiency", proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services, Munich, Germany. Website: https://doi.org/10.1145/2493190.2493204

Saket, Bahador, Endert, Alex, and Stasko, John. (2016). "Beyond Usability and Performance: A Review of User Experience-Focused Evaluations in Visualization", proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV' 16). Website: https://doi.org/10.1145/2993901.2993903

Thillaieaswaran, B. and Pasupathy, S. (2021). Learnability Metric in Software Quality Assurance, Annals of the Romanian Society for Cell Biology Volume 25 No. 2. Website: https://www.annalsofrscb.ro/index.php/journal/article/view/1406

World Economic Forum. (2023). What is Generative AI? An AI Explains Website: https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/