Investigating the Effect of Sample Size and Respondent Characteristics on Usability Measurement: The Case of ChatGPT

Jakub Swacha
University of Szczecin
Szczecin, Poland *jakub.swacha@usz.edu.pl*

Łukasz Radliński
West Pomeranian University of Technology in Szczecin
Szczecin, Poland *lukasz.radlinski@zut.edu.pl*

Karolina Muszyńska
University of Szczecin
Szczecin, Poland *karolina.muszynska@usz.edu.pl*

Susanne Marx
Technical University of Applied Sciences Wildau
Wildau, Germany *susanne.marx@th-wildau.de*

Ricardo Queirós
Polytechnic of Porto
Porto, Portugal *ricardoqueiros@esmad.ipp.pt*

Abstract

Standardized usability questionnaires are a fast and relatively effortless way of assessing usability of software products. Despite their long use, so far, little attention has been paid to the effect of sample size and the level of respondents' acquaintance with the evaluated software on the measurement. This paper addresses this gap and uses SUS and mTAM measurements of ChatGPT to illustrate how the deviation from the mean usability score decreases with increasing sample size, and to confirm the significant effect of usage frequency and knowledge of the evaluated software and its alternatives on the measurement results. It also exposes no demographic bias due to participation of respondents of different gender, country of stay, and academic major.

Keywords: ChatGPT, modified Technology Acceptance Model, standardized usability questionnaire, System Usability Scale, sample size.

1. Introduction

The term *usability* came into general use in the early 1980's. Although the goal of usability engineering has evolved since then from developing products that are objectively effective, efficient, and which will make their users satisfied to a wider notion of user experience (UX), which, includes attention to emotional factors such as pleasure, beauty, and trust, usability remains at the core of human–computer interaction research [21, p. 973].

Usability is counted among the human-focused factors that are critical for the success of digital transformation endeavours [8, p. 593][27, p. 7]. Poor usability creates barriers for digital transformation of businesses and industries [34, p. 137], therefore monitoring usability is an essential element of digital transformation which strives to consider both technological and human-centered aspects [7, p. 2]. As usability cannot be measured directly, and what can be measured are its certain aspects, a plethora of methods and measures have been exploited to capture them [12, p. 80].

Such a variety of available methods brings several issues concerning research objectivity, replicability, quantification, economy, communication, and scientific generalization, which can be addressed by standardization of usability measures [29, p. 185–186]. Consequently, a notable effort has been made in this direction, of which the most evident product is the standardized questionnaires intended to assess perceived usability [2, p. 15]. While it can be argued that such self-administered questionnaires are insufficient to effectively measure all aspects of usability defined in ISO 9241–11 (in particular, effectiveness and efficiency), this can be addressed by complementing their subjective results with objective measurements (of, e.g., the number of user's keystrokes executing a given task, the amount of time the user spent on productive versus unproductive actions while completing tasks, and the user's task success rate) [32, p. 4–5].

Standardized does not mean commonly agreed, and already in 2023, there were 27 usability questionnaires in wide use [10, p. 138]. They differ in scope (from 2 usability aspects covered by UMUX-Lite to 16 by SUMI [2, p. 24]), the number of question items (from only 1 in SEQ to 100 in PUTQ [2, p. 17]), and the used measurement scale (from 3-point dichotomous scale in SUMI to a graduated scale from 0 to 150 in SMEQ [2, p. 17]). While several works have been devoted to the investigation of correspondence between these questionnaires in various contexts [4, 5], [13], [16], [19], [35], or the effect of various modifications of the format and order of questionnaire items and responses [14], [18], [20], [28], notably less research interest has been put on how the choice of sample size and respondents' characteristics affect usability measurement with standardized questionnaires [25], [33], [26], [23].

This paper aims to address this gap by investigating the effect on the results of measurement using two standardized usability questionnaires caused by a difference in the size of the sample, frequency of using the assessed system, user's knowledge of the assessed system and its alternatives. Three research questions are stated:
– RQ1: What is the size of deviation of usability scores based on small samples from those based on large samples and whether it is negatively or positively biased?
– RQ2: What is the effect of respondents' characteristics related to their level of acquaintance with the software on the measured usability score?
– RQ3: Do the general demographic traits significantly affect usability scores?

The presented study also contributes practically to the pool of usability measurements by providing a new set of results for ChatGPT which has recently become a very popular software.

2. Related Work

2.1. Standardized Usability Questionnaires: Definitions and Adaptations

Among the 27 standardized usability questionnaires that attained some popularity so far [10, p. 138], for the sake of keeping the total survey time in limits, two relatively short popular questionnaires were selected for the research presented here.

The first of them is System Usability Scale (SUS), originally developed by John Brooke at Digital Equipment Corporation in 1986, which over the years, has become an instrument commonly utilized in usability testing of various kinds of products [11]. The SUS questionnaire, originally in English, has been translated to many other languages, including: Arabic, Chinese, Danish, French, German, Hindi, Indonesian, Italian, Polish, Portuguese, Malay, Slovene, Spanish, and Turkish [35, p. 2 and works cited therein].

SUS comprises 10 items measured using a 5-point Likert scale, ranging from "strongly disagree" to "strongly agree". Its original version [6], in an attempt to reduce the acquiescent response bias induced by respondents who tend to agree with all or almost all statements in a questionnaire, included a mix of positively and negatively worded items (see the left side of Table 1). Although, in theory, this should force attentive respondents to disagree with some statements as well as help to identify (and remove from the sample) serial extreme responders,

i.e., participants who provide all high or all low ratings, in practice, it may create more problems than it solves due to: misinterpretation, as users may respond differently to negatively worded items such that reversing their scores does not account for the difference, which leads to creating an artificial two-factor structure and lowering internal reliability; mistakes of users who may forget a statement is negative and accidentally agree with it when they meant to disagree; and miscoding by researchers who may forget to reverse the scales when scoring, and consequently report incorrect data [28]. To overcome these problems, Sauro and Lewis proposed an all-positive version of the SUS questionnaire [28] (see the right side of Table 1).

For the sake of comparison between different usability measurements, SUS yields a single number from a range of 0 to 100, representing a composite measure of the overall usability of the system being assessed [6]. The SUS score is the sum of score contributions from each item multiplied by 2.5. The score contribution of each positive item is the scale position minus 1, whereas for each negative item (in the original SUS version), it is 5 minus the scale position.

Kortum et al., who analyzed the SUS measurements of 20 products, found that both the standard and all-positive versions of SUS can be used confidently to measure subjective usability and that the scores generated from both SUS versions can be directly compared [14].

Table 1. Original SUS vs All-positive SUS.

#Item	Original SUS [6]	All-positive SUS [28]
1.	I think that I would like to use this system frequently.	I think that I would like to use this system frequently.
2.	I found the system unnecessarily complex.	I found this system to be simple.
3.	I thought the system was easy to use.	I thought the system was easy to use.
4.	I think that I would need the support of a technical person to be able to use this system.	I think I could use this system without the support of a technical person.
5.	I found the various functions in this system were well integrated.	I found the various functions in this system were well integrated.
6.	I thought there was too much inconsistency in this system.	I thought there a lot of consistency in this system.
7.	I would imagine that most people would learn to use this system very quickly.	I would imagine that most people would learn to use this system very quickly.
8.	I found the system very cumbersome to use.	I found the system very intuitive.
9.	I felt very confident using the system.	I felt very confident using the system.
10.	I needed to learn a lot of things before I could get going with this system.	I could use the system without having to learn anything new.

Although a 100-point scale used for the SUS score is easy to understand and allows for relative judgments, it does not provide information on how the numeric score translates into an absolute judgment of usability. To overcome this gap, Bangor et al. proposed an adjective rating scale to help interpret individual SUS scores and explain the results [3, p. 114]. It associates the adjective "worst imaginable" with the mean SUS score of 12.5, "awful" with 20.3, "poor" with 35.7, "OK" with 50.9, "good" with 71.4, "excellent" with 85.5, and "best imaginable" with 90.9 [3, p. 118].

The same authors also proposed a letter grade scale modeled after the American school grading system, with "A" given for the mean SUS score of 90 or higher, "B" for the SUS score between 80 and 89, "C": 70–79, "D": 60–69, "E": 50–59, and F for any SUS score below 50 [3, p. 120–121]. It was, however, criticized by Sauro and Lewis who showed that it was virtually impossible to get an "A" grade according to it, and suggested an alternative, curved grading scale, in which "A+" grade is given for the SUS score of over 84, "A" for the SUS score in the range of 80.8–84.0, "A-": 78.9–80.7, "B+": 77.2–78.8, "B": 74.1–77.1, "B-": 72.6–74.0, "C+":

71.1–72.5, "C": 65.0–71.0, "C-": 62.7–64.9, "D": 51.7–62.6, and "F" for the SUS score of 51.6 or less [29, p. 203–204].

The second of the chosen questionnaires, mTAM [16], is a modified version of the questionnaire developed by Davis to measure two fundamental constructs of his Technology Acceptance Model (TAM): Perceived Usefulness (PU; users' belief that using the technology would improve their performance) and Perceived Ease of Use (PEU; users' belief that using the technology would not require significant effort) [9].

Table 2. TAM vs mTAM.

#Item	TAM [9]	mTAM [16]
	Perceived Usefulness	
1.	Using this system in my job would enable me to accomplish tasks more quickly.	Using this system enabled me to accomplish tasks more quickly.
2.	Using this system would improve my job performance.	Using this system improved my job performance.
3.	Using this system in my job would increase my productivity.	Using this system in my job increased my productivity.
4.	Using this system would enhance my effectiveness on the job.	Using this system enhanced my effectiveness on the job.
5.	Using this system would make it easier to do my job.	Using this system made it easier to do my job.
6.	I would find this system useful in my job.	I found this system useful.
	Perceived Ease of Use	
1.	Learning to operate this system would be easy for me.	It was easy to learn to operate this system.
2.	I would find it easy to get this system to do what I want it to do.	I found it easy to get this system to do what I wanted it to do.
3.	My interaction with this system would be clear and understandable.	My interaction with this system was clear and understandable.
4.	I would find this system to be flexible to interact with.	I found this system to be flexible to interact with.
5.	It would be easy for me to become skillful at using this system.	It was easy for me to become skillful at using this system.
6.	I would find this system easy to use.	I found this system easy to use.

The TAM questionnaire comprises two sets of six items (for PU and PEU, respectively), measured with a 7-point semantic scale, ranging from "extremely likely" to "extremely unlikely" (see the left side of Table 2). Although its original purpose was to explain and predict a person's acceptance and adoption of a specific system, it has become one of the popular instruments used for usability measurement [2]. To make the TAM questionnaire more suitable for this new role, Lah et al. proposed to modify the orientation of some of its item statements from the imagined future to the past, as well as adopting a 7-point Likert scale (also reverting the original order to the one used in other standardized usability questionnaires, i.e., from "strongly disagree" to "strongly agree"), calling the modified questionnaire mTAM [16, p. 5] (see the right side of Table 2). According to the results from the same source, mTAM is characterized by high reliability (respectively, 0.95, 0.95, and 0.97 in the three performed surveys, assessing Gmail, PowerPoint, and IBM Notes) and strong correlation with SUS (0.80, 0.70, and 0.90) [16, p. 5–8]. Wang and Wang reported similar results from their survey on Tencent Meeting mobile application performed with Chinese versions of the questionnaires (mTAM reliability: 0.93, correlation with SUS: 0.83) [35].

2.2. Prior Research on the Effect of Sample Size and Respondent Characteristics on Standardized Usability Questionnaire Results

There is a huge discrepancy in the reported number of respondents involved in the surveys based on standardized usability questionnaires. Among the sources quoted in Table 3, it spanned from 10 in [36] to 194 in [1], with an average of 93 respondents. Regarding the minimum sample size needed to get reliable results, many papers refer to the number of 12, taken from the study of Tullis and Stetson who randomly retrieved subsamples of increasing size from 6 to 14 from a pool of 123 SUS results for two websites: finance.yahoo.com and kiplinger.com, and found out that already a subsample of size 12 was sufficient to correctly indicate the better website of the two [33]. At the moment of writing these words, Google Scholar reports 1208 citations of that study, indicating its notable popularity, and there was at least one successful attempt at replicating these results with regard to two learning management systems (eClass and Moodle) [25]. Nonetheless, we cannot consider the number 12 as sufficient for obtaining reliable results from surveys using standardized usability questionnaires as that study has not shown that a sample of 12 would give the same SUS score as a sample of 123, only that it was sufficient to indicate one of two websites with a higher score consistently with the sample of 123; moreover, the SUS scores of those two websites were vastly different (50 vs. 73 [33, Fig. 6 therein]), so a sample of 12 would probably be far from sufficient if the difference was subtle.

McLellan et al. investigated the effect of experience on SUS score of two related software products among 262 end users across different geographic locations, reporting that users having a more extensive experience with the assessed software provided as much as 15-16% higher SUS scores than users with lesser experience, regardless of the software product [23].

Robertson and Kortum investigated the effect of cognitive fatigue on SUS results on a sample of 43 participants using twelve prototype paper voting ballots, with each participant having to complete six ballots before a fatigue manipulation and six after it, finding no significant difference in the SUS score measured pre- and post-fatigue [26].

Lorenz et al. quantified the effects of age and gender on several metrics, including SUS score, in a study involving 57 users of a Virtual-Reality mobile game, finding that younger users provided significantly higher SUS scores. Yet no difference between genders nor any interaction effects of age and gender [22].

2.3. Previous Reports on ChatGPT Usability Assessment

Using Web of Science and ScienceDirect databases, amended by a search on Google Scholar, revealed numerous papers reporting SUS scores for ChatGPT in various contexts. Most of the authors used the original SUS scale [1], [17], [24], [30, 31], [36], whereas one applied UMUX-Lite [15] yet recalculated the results to SUS score. Their comparison, including the source, ChatGPT usage scope or target user group, the number of respondents, the obtained SUS score, and SUS reliability measured with Cronbach's α (where provided) are shown in Table 3.

Table 3. Previously published ChatGPT usability assessments.

Source	Scope	Country	Users	SUS score	Reliability
[1]	Clinical questions	Saudi Arabia	194	64.52	0.76
[15]	Physics questions	Germany	27	73.05	–
[17]	Academic writing	Finland	50	76.25	–
[24]	General (University students)	Indonesia	121	67.44	0.61
[30]	Clinical questions	Germany	40	83.38	0.72
[31]	Qualitative data analysis	Slovenia	85	79.03	0.85
[36]	Environmental education	China	10	87.50	–

3. Research Method and Data Collection

Four European academic institutions were approached to collect usability measurements needed for the analysis: University of Szczecin and West Pomeranian University of Technology (both in Poland), Polytechnic of Porto (Portugal), and Technical University of Applied Sciences Wildau (Germany). In each institution, the data were collected anonymously using an online questionnaire consisting of three parts: (1) user context data: Usage frequency of this system, Knowledge of the system, Knowledge of the system alternatives, Academic major, Gender, Country of education; (2) the all-positive version of the SUS questionnaire as defined in [28]; (3) the modified version of the TAM questionnaire (mTAM) as defined in [16].

To avoid confusion among respondents who had to answer two questionnaires one after another, we decided to consistently use a 5-point Likert scale also for mTAM (instead of the originally suggested 7-point). According to the study by Lewis [20], such a change in the number of response options should not affect the results significantly.

The students were asked to fill in the questionnaire twice – the second time not earlier than a few hours and not later than a week after providing initial responses. An additional questionnaire form field ("magic number") was used to pair the responses provided by the same respondent. The responses were collected between 14 and 30 April 2025 (various groups were surveyed on different dates). In total, 444 responses have been collected from 259 respondents, including 171 respondents who provided at least two responses.

The purpose of the repeated survey was to ensure respondents' reliability and thus, a satisfactory data quality. The responses from respondents who provided vastly contrasting assessments in their two responses were deleted, i.e., with a mean score difference exceeding 1 point per questionnaire item – which could not be attributed to a change in user experience, considering a very short time between the first and the repeated survey. Data cleaning also included the removal of responses from respondents who filled in their questionnaires unreasonably quickly (in less than 40 seconds) as well as those who selected the answer "I have never seen ChatGPT or heard about it" to the "Knowledge of ChatGPT" questionnaire item or selected the answer "I have never used ChatGPT" to the "Usage frequency of ChatGPT" questionnaire item – as they cannot know about the ChatGPT usability.

After cleaning and leaving only the second response from each pair of responses of the same respondent (assuming it to be more deliberate as the respondent had time for second thought), the analyzed dataset contained 231 responses, including 182 from males and 46 from females (3 persons selected the "Other/Don't want to answer" option). The academic majors for respondents were computer science (192 students), information technology and econometrics (22), management (15), and digital health (2). The respondents studied at the four universities in three countries, in particular: in Poland (184) – West Pomeranian University of Technology in Szczecin (164) and University of Szczecin (22), in Portugal – Polytechnic of Porto (32), and in Germany – Technical University of Applied Sciences Wildau (15).

The reliability measured with Cronbach's alpha was 0.868 for SUS and 0.901 for mTAM (as for its subdimensions, respectively, 0.863 for PU and 0.876 for PEU). All these values are well above the widely-agreed threshold of 0.7.

4. Results

4.1. Effect of Increasing Sample Size

Taking into consideration all responses that passed the cleaning stage (231 in total), the resulting SUS score is 75.67 and mTAM score is 79.18. The former result translates to good rating according to Bangor et al. [3, p. 114] or grade "B" according to Sauro and Lewis [29, p. 203–204].

To address RQ1, for both SUS and mTAM scores, we have compared the minimum and

maximum mean that could be obtained for x samples retrieved from the analyzed set, with x increasing from 3 to 231. The results are depicted in Fig. 1.

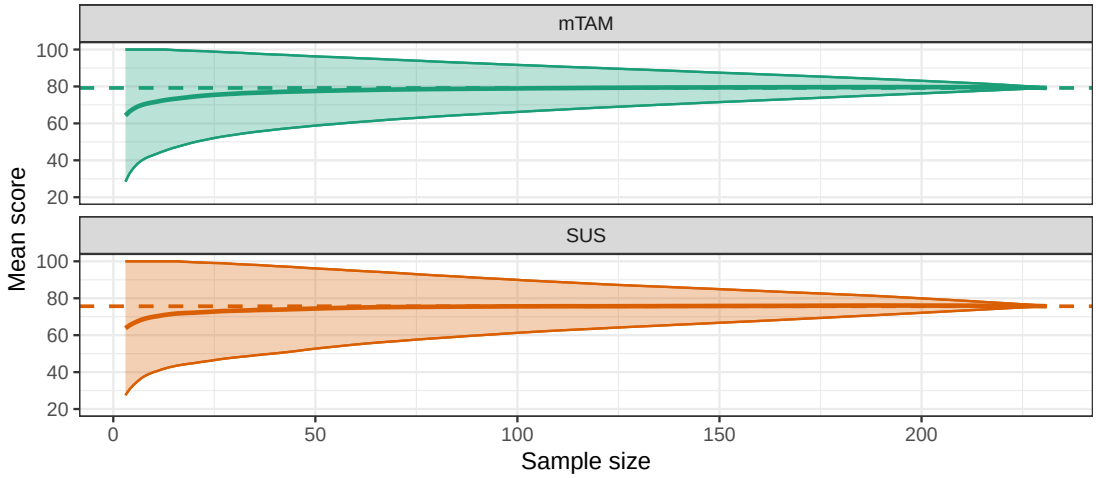

Fig. 1. Impact of the sample size on UX scores.

As shown in Fig. 1, the usability measurements obtained with very small samples can be very volatile. Moreover, as indicated by the dark solid line representing the mean of the minimum and maximum means for a given sample size (in distinction to the dark dashed line showing the mean calculated for the whole analyzed set), the results obtained for small sample sizes can be systematically underestimated, as, seemingly, the respondents are more apt to provide extremely negative assessments than extremely positive ones, and the effect of extreme assessments on the mean score disappears only with larger samples. To provide more details on these findings, Table 4 shows the maximum and minimum scores that could be obtained from the analyzed set using the specified sample size as well as the average absolute deviation (AAD) from the mean score (of the whole analyzed dataset) for a given sample size estimated using 10,000 random samples of the specified size.

Table 4. The maximum under- and overestimate in percents, and the estimated average absolute deviation from the mean UX score for a given sample size.

Sample size	SUS			mTAM		
	underestimate%	AAD	overestimate%	underestimate%	AAD	overestimate%
3	63.66	7.36	32.15	64.04	6.48	26.29
5	56.39	5.55	32.15	54.75	5.04	26.29
10	47.14	3.97	32.15	45.80	3.57	26.29
14	43.60	3.29	32.15	41.74	3.03	26.10
25	38.55	2.52	30.96	34.23	2.27	24.81
50	30.36	1.79	27.06	25.75	1.60	21.60
100	19.02	1.27	18.87	16.41	1.13	15.76
200	4.67	0.90	5.64	3.73	0.80	4.90

As can be observed in Table 4, measurements obtained from small samples can be very distant from those based on large samples. With only five respondents (which is the number of people often recruited for usability tests), the extreme possible SUS scores obtained from the analyzed dataset could be over 50% higher or 30% lower than the actual mean from all 234 samples. Of course, the probability of obtaining so highly distorted results is very low,

nonetheless, on average, the score based on five measurements differs by ± 5 points from the mean which can easily put the result in a different interval of the curved grading scale [29, p. 203–204]. The estimated average absolute deviation drops below 2.5 points at 26 samples; with 100 samples, it is below 1.3. The bias due to the predominance of the maximum underestimate over the overestimate becomes negligible also at around 100 samples. The presented results also indicate that mTAM is relatively less prone to distorted measurements using small sample sizes than SUS.

4.2. Impact of Users' Knowledge and Usage of the Assessed Software

There were six respondent characteristics described by data collected in the first section of the survey, three of them describing general demographic traits (Academic major, Gender, Country of education) and three connected to the level of user's acquaintance with the software (Usage frequency and Knowledge of the software under assessment, knowledge of Alternative software of similar kind). Table 5 illustrates the relationships between these attributes and SUS and mTAM scores. Depending on the attribute type, this analysis involved diverse statistical tests and effect size measures. Specifically, we used Spearman's rank correlation coefficient ρ for the first three ordinal attributes (both as a statistical test and effect size measure), Mann-Whitney's test and Cliff's delta for Gender, Kruskal-Wallis test and epsilon-squared for the last two nominal attributes. As mentioned earlier, three respondents provided answers for Gender other than male/female. Since such a low number of cases was too low for a separate category and might bias calculated relationships, we removed these three cases from this analysis. For the same reason, we removed two cases when analyzing the Academic major. The p-values were adjusted using Bonferroni correction to control the family-wise error rates. To determine statistical significance, we used the $\alpha = 0.05$.

Usage frequency, Knowledge of ChatGPT, and Knowledge of alternatives were all statistically significantly correlated with both SUS and mTAM. Only for Usage frequency, these correlations were moderate in strength, while for Knowledge of ChatGPT and Knowledge of alternatives, they were weak. Correlations with mTAM were slightly higher than with SUS. All these correlations were positive, i.e., with the increasing level for each attribute, the value of SUS and mTAM also increased. Additional details, i.e., the mean mTAM and SUS scores grouped by particular categories of Usage frequency, Knowledge of ChatGPT, and Knowledge of alternatives, were illustrated in Fig. 2. Note that this figure includes cases from respondents who never used ChatGPT and have no knowledge of it. After omitting them, the results both in Table 5 and Fig. 2 show that respondents who frequently used ChatGPT, knew it better, and knew its alternatives were also more satisfied with such use.

All relationships between Gender, Academic major, and Country of stay with both SUS and mTAM were not statistically significant, while the strength of effect sizes was small. Such results indicate that SUS and mTAM scores did not meaningfully vary by these demographics. Thus, they demonstrate the lack of demographic bias.

5. Discussion

Regarding RQ1, we have demonstrated how the sample size affects maximum measurement errors and the average absolute deviation from the mean score (of the whole analyzed dataset), which indicates that the measurement based on very small sample sizes (counting 5 respondents or less) is highly unreliable as the average deviation then exceeds the half of the curved grade scale interval length. The deviation for the sample size of 14, recommended in [33], is still over 3 points. In our opinion, such a sample size could therefore be sufficient if the same group of respondents compares the usability of different versions of the same software or alternative software products serving the same purpose. For usability benchmarking purposes, we recommend

Table 5. Relationships of respondent data with UX scores*.

Attribute	N	SUS		mTAM	
		Statistic and effect size	***p* adj.**	**Statistic and effect size**	***p* adj.**
Usage frequency	231	$\rho = 0.410$	< 0.001	$\rho = 0.501$	< 0.001
Knowledge of ChatGPT	231	$\rho = 0.279$	< 0.001	$\rho = 0.380$	< 0.001
Knowledge of alternatives	231	$\rho = 0.208$	0.018	$\rho = 0.313$	< 0.001
Gender	228	W = 4,994.5, $\delta = 0.193$	0.515	W = 4,739.5, $\delta = 0.132$	1.000
Academic major	229	H = 0.605, $\epsilon^2 = 0.003$	1.000	H = 0.826, $\epsilon^2 = 0.006$	1.000
Country of stay	231	H = 0.511, $\epsilon^2 = 0.002$	1.000	H = 1.284, $\epsilon^2 = 0.006$	1.000

* ρ – Spearman's rank correlation coefficient, W – Mann-Whitney's test W, H – Kruskal-Wallis test H, δ – Cliff's delta, ϵ^2 – epsilon-squared.

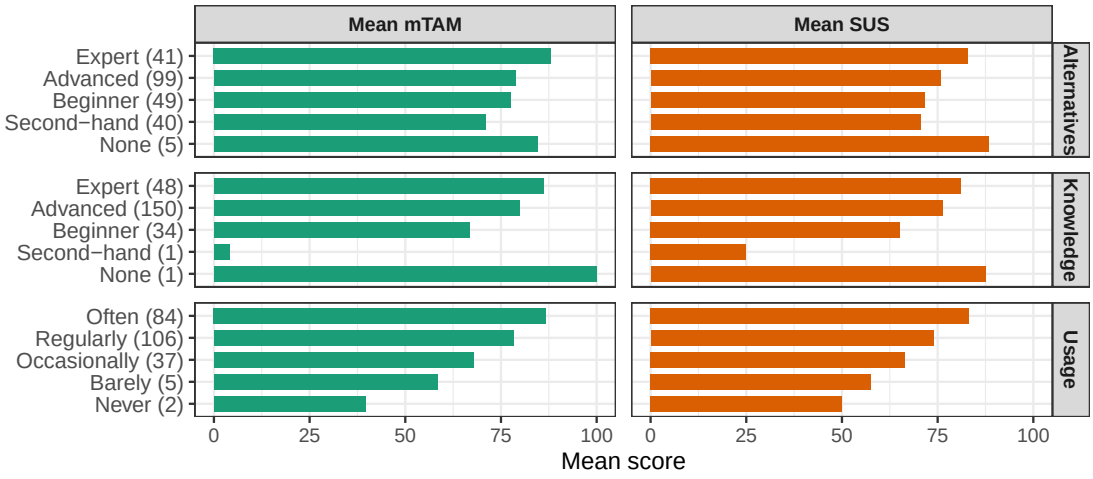

Fig. 2. Categorized scores. The values in brackets indicate case counts.

to survey groups of at least 100 respondents.

Regarding RQ2, we have confirmed the significant effect of respondents' characteristics related to their level of acquaintance with the software on the measured usability score. As regards the strength of the effect, it was moderate for Usage frequency and small for both Knowledge of the evaluated software and Knowledge of its alternatives. In all these cases, it was positive. Regarding the knowledge of the evaluated software, this confirms the prior findings of McLellan et al. [23], though the effect we measured was almost twice as strong. Regarding the two other variables (Usage frequency and Knowledge of alternatives), our study is the first to consider them, so establishing their significant effect is an important finding for future research on usability measurement.

Regarding RQ3, we have identified no statistically significant effect of respondents' gender, country or academic major on both SUS and mTAM scores (and the strength of the effect was negligible in all three cases). The result regarding gender is consistent with that of Lorenz et al. [22]. Overall, this is also an important finding as it negates the presence of demographic bias, therefore suggesting that usability measurements obtained from respondents of different gender, from various countries, and having different educational backgrounds can be combined.

Regarding the ChatGPT usability measurements, the measured SUS score of 75.67 is closely between the values reported earlier in [15] (73.05) and [17] (76.25). The larger distance to other results reported in the literature, e.g., 67.44 by Mulia et al. who also surveyed university students [24] could be, at least in part, attributed to other factors, possibly those that have been identified in this study as relevant (e.g., usage frequency) but were not measured by Mulia et al.

6. Threats to Validity

We used well-established instruments (all-positive SUS [28] and mTAM [16]) for comparability with prior work on ChatGPT (e.g., [1], [24], [31]. High reliability (Cronbach's alpha: 0.868 for SUS, 0.901 for mTAM) suggests internal consistency. These standardized questionnaires allow capturing subjective perceptions, but were designed for software interfaces that were more static or traditional. Therefore, they might not fully reflect the characteristics of generative AI tools like ChatGPT, which are highly dynamic and context-dependent, in which the answers may change in different sessions. We have not formally re-validated SUS/mTAM specifically for generative AI. Future work could complement standardized questionnaires with task-based objective measures (e.g., success rates on designed prompts), qualitative feedback, or LLM-specific scales that address dynamic response quality, consistency, and trust.

The questionnaire presented questions for each respondent in a fixed order, i.e., user-context first, then mTAM, and then SUS. This could introduce priming or fatigue: responding to mTAM items may influence subsequent SUS responses. We kept the order consistent across all participants, minimizing noise from varying the order of questions.

Online surveys can be affected by respondents feeling pressure to provide satisfying or socially desirable answers. To tackle this, surveys were conducted anonymously, reducing such pressure. We also removed responses completed implausibly quickly (< 40s) to mitigate low-engagement responses. The repeated-measures design (two questionnaire sessions) and removal of inconsistent pairs further guarded against careless or disengaged responses. In future, embedding attention checks or including objective task performance measures could further guard against response bias.

With 231 cleaned responses, performed analyses had adequate power to detect moderate effects (e.g., usage frequency). We set up a significance level $\alpha = 0.05$ and applied the Bonferroni correction for multiple tests. Bootstrap analyses (Fig. 1, Table 4) confirmed the stability of mean estimates with this size. Although the overall sample is sufficient, subgroup analyses (e.g., by gender or country) had smaller cell sizes; non-significant results there may reflect low power for small subgroups.

All participants were students, likely with relatively high digital fluency and specific use contexts (e.g., academic tasks). Obtained findings may not generalize to other populations, such as professionals in different domains, older adults, or non-academic users. We recruited participants from four institutions across three countries and four academic majors to introduce diversity in the educational context and cultural background. Nevertheless, all were higher-education students. Future mitigation may include non-student respondents, older demographics, and varied cultural and educational backgrounds.

Invitations were sent broadly to whole student groups to reduce self-selection bias, but participation remained voluntary. We acknowledge possible bias related to respondent participation as respondents with stronger opinions (positive or negative) or greater interest in ChatGPT might have been more likely to respond.

We did not restrict participants to specific tasks with ChatGPT. Therefore, respondents evaluated the system based on their diverse personal uses. Such an approach reduced control over task context, which could affect perceived usability. We partially accounted for varied experience levels by collecting data on Usage frequency and Knowledge of alternatives. However, this did not isolate context effects.

7. Conclusion

Despite the popularity of standardized usability questionnaires, so far there was no research which would seriously investigate to what extent the usability measurement can be distorted by performing it on a sample of small size, or how much the measurement result is affected by the frequency of use of the assessed system by the respondents, their knowledge of it, and their knowledge of its alternatives. The presented study provides answers to all these questions. Its main research contributions are (1) showing the correspondence between sample size and measurement deviation from the mean calculated on a large sample which helps to choose sample sizes adequate to the desired level of measurement precision; (2) showing the difference in usability measurement between respondents having various levels of knowledge and experience in the use of the evaluated software which suggests to choose users at various proficiency levels to gain more realistic assessment of software usability.

These results will come in handy for designing and performing usability assessment in digital transformation initiatives, helping to attain reliable measurements by choosing a more adequate size and composition of the evaluation group. Moreover, the presented results have also extended the existing pool of ChatGPT usability measurements, enabling future comparisons.

Acknowledgements

The presented work has been co-financed by the Minister of Science of Poland under the Regional Excellence Initiative program.

References

[1] Aljamaan, F., Malki, K.H., Alhasan, K., Jamal, A., Altamimi, I., Khayat, A., Alhaboob, A., Abdulmajeed, N., Alshahrani, F.S., Saad, K., Al-Eyadhy, A., Al-Tawfiq, J.A., Temsah, M.H.: ChatGPT-3.5 System Usability Scale early assessment among healthcare workers: Horizons of adoption in medical practice. Heliyon 10(7) (2024)

[2] Assila, A., Marçal de Oliveira, K., Ezzedine, H.: Standardized usability questionnaires: Features and quality focus. Electronic Journal of Computer Science and Information Technology 6(1) (2016)

[3] Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of Usability Studies 4(3), pp. 114–123 (2009)

[4] Berkman, M.I., Karahoca, D.: Re-assessing the usability metric for user experience (UMUX) scale. Journal of Usability Studies 11(3), pp. 89–109 (May 2016)

[5] Borkowska, A., Jach, K.: Pre-testing of Polish Translation of System Usability Scale (SUS). In: Proceedings of 37th International Conference on Information Systems Architecture and Technology – ISAT 2016 – Part I. pp. 143–153. Springer, Cham (2017)

[6] Brooke, J.: SUS - A quick and dirty usability scale. In: Jordan, P., Thomas, B., Weerdmeester, B., McClelland, I. (eds.) Usability evaluation in industry, pp. 189–194. Taylor & Francis, London (1996)

[7] Chrusciak, C.B., Szejka, A.L., Canciglieri Junior, O.: Integrating digital transformation with human-centric factors strategies to enhance organisational process performance: The H.O.P.E. model. Journal of Industrial Information Integration 44 (2025)

[8] Chrusciak, C.B., Szejka, A.L., Schaefer, J.L., Canciglieri Junior, O.: Towards a Digital Transformation and Human Factors Integrated Framework: Application of Structural

Equation Modelling. In: Cooper, A., Trigos, F., Stjepandić, J., Curran, R., Lazar, I. (eds.) Engineering for Social Change. IOS Press (2024)

[9] Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly pp. 319–340 (1989)

[10] Graser, S., Böhm, S.: Quantifying user experience through self-reporting questionnaires: A systematic analysis of the sentence similarity between the items of the measurement approaches. In: HCI International 2023 – Late Breaking Posters. pp. 138–145. Springer Nature, Cham (2024)

[11] Grier, R.A., Bangor, A., Kortum, P., Peres, S.C.: The system usability scale: Beyond standard usability testing. In: Human Factors and Ergonomics Society annual meeting. vol. 57, pp. 187–191. Sage, Los Angeles (2013), issue: 1

[12] Hornbæk, K.: Current practice in measuring usability: Challenges to usability studies and research. Int'l J. of Human-Computer Studies 64(2), pp. 79–102 (Feb 2006)

[13] Kocabalil, A.B., Laranjo, L., Coiera, E.: Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In: Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI) (2018)

[14] Kortum, P., Acemyan, C.Z., Oswald, F.L.: Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). Human Factors 63(6), pp. 987–998 (2021)

[15] Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K.E., Lukowicz, P., Kuhn, J., Küchemann, S., Karolus, J.: Unreflected acceptance - investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In: HHAI 2024: Hybrid Human AI Systems for the Social Good. IOS Press (2024)

[16] Lah, U., Lewis, J.R., Šumak, B.: Perceived Usability and the Modified Technology Acceptance Model. Int'l J. of Human–Computer Interaction 36(13), pp. 1216–1230 (2020)

[17] Launonen, P., Talalakina, E., Dubova, G.: Students' Perceptions of Using ChatGPT for Academic Writing in English: Insights from a Finnish University. Tertium Linguistic Journal 9(1), pp. 219–249 (2024)

[18] Lewis, J.R.: Comparison of Four TAM Item Formats: Effect of Response Option Labels and Order. Journal of Usability Studies 14(4) (2019)

[19] Lewis, J.R.: Measuring Perceived Usability: SUS, UMUX, and CSUQ Ratings for Four Everyday Products. Int'l J. of Human–Computer Interaction 35(15), pp. 1404–1419 (2019)

[20] Lewis, J.R.: Measuring User Experience With 3, 5, 7, or 11 Points: Does It Matter? Human Factors 63(6), pp. 999–1011 (Sep 2021)

[21] Lewis, J.R., Sauro, J.: Usability and User Experience: Design and Evaluation. In: Salvendy, G., Karwowski, W. (eds.) Handbook of Human Factors and Ergonomics, pp. 972–1015. Wiley (2021)

[22] Lorenz, M., Brade, J., Klimant, P., Heyde, C.E., Hammer, N.: Age and gender effects on presence, user experience and usability in virtual environments–first insights. Plos One 18(3), pp. 1–16 (03 2023)

[23] McLellan, S., Muddimer, A., Peres, S.C.: The effect of experience on System Usability Scale ratings. Journal of Usability Studies 7(2), pp. 56–67 (2012)

[24] Mulia, A.P., Piri, P.R., Tho, C.: Usability analysis of text generation by ChatGPT OpenAI using System Usability Scale method. Procedia Comp. Sci. 227, pp. 381–388 (2023)

[25] Orfanou, K., Tselios, N., Katsanos, C.: Perceived usability evaluation of learning management systems: Empirical evaluation of the System Usability Scale. The International Review of Research in Open and Distributed Learning 16(2) (2015)

[26] Robertson, I., Kortum, P.: The Effect of Cognitive Fatigue on Subjective Usability Scores. Proc. of the Human Factors and Ergonomics Society Annual Meeting 61(1), pp. 1461–1465 (2017)

[27] Sahu, N., Deng, H., Mollah, A.: Investigating the critical success factors of digital transformation for improving customer experience. In: Proc. of the International Conference on Information Resources Management (CONF-IRM). AIS, Melbourne (2018)

[28] Sauro, J., Lewis, J.R.: When designing usability questionnaires, does it hurt to be positive? In: SIGCHI Conference on Human Factors in Computing Systems. pp. 2215–2224. ACM, New York (2011)

[29] Sauro, J., Lewis, J.R.: Standardized usability questionnaires. In: Quantifying the User Experience. Practical Statistics for User Research, pp. 185–248. Elsevier, Amsterdam (2016)

[30] Stroop, A., Stroop, T., Zawy Alsofy, S., Wegner, M., Nakamura, M., Stroop, R.: Assessing GPT-4's accuracy in answering clinical pharmacological questions on pain therapy. British Journal of Clinical Pharmacology (2025)

[31] Šumak, B., Pušnik, M., Kožuh, I., Šorgo, A., Brdnik, S.: Differences in user perception of artificial intelligence-driven chatbots and traditional tools in Qualitative Data Analysis. Applied Sciences 15(2: 631) (2025)

[32] von Sydow, T.: Standardizing Usability Evaluation: Case study of objective measures as complements to user satisfaction surveys. Master's thesis, Uppsala University, Department of Informatics and Media (2022)

[33] Tullis, T.S., Stetson, J.N.: A comparison of questionnaires for assessing website usability. In: Usability professional association conference. vol. 1, pp. 1–12. Usability Professional Association, Minneapolis (2004)

[34] Uzule, K., Verina, N.: Digital Barriers in Digital Transition and Digital Transformation: Literature Review. Economics and Culture 20(1), pp. 125–143 (2023)

[35] Wang, Y., Wang, H.: Measuring Perceived Usability in Chinese Questionnaires: mTAM, SUS, and UMUX. Int'l J. of Human–Computer Interaction 38(11), pp. 1052–1063 (2022)

[36] Xiaoyu, W., Zainuddin, Z., Leng, C.H., Wenting, D., Li, X.: Evaluating the efficacy of ChatGPT in environmental education: findings from heuristic and usability assessments. On the Horizon: The International Journal of Learning Futures 33(2), pp. 165–185 (2025)