A Web-Based Analysis Toolkit for the System Usability Scale

Jonas Blattgerste*
jonas.blattgerste@hs-emden-leer.de
University of Applied Sciences Emden/Leer
Emden, Lower Saxony, Germany

Jan Behrends*
jan.behrend@stud.hs-emden-leer.de
University of Applied Sciences Emden/Leer
Emden, Lower Saxony, Germany

Thies Pfeiffer
thies.pfeiffer@hs-emden-leer.de
University of Applied Sciences Emden/Leer
Emden, Lower Saxony, Germany

ABSTRACT

The System Usability Scale (SUS) questionnaire is a broadly used usability measurement tool, which is fast in its application and straight forward in its interpretation. While the original SUS questionnaire was envisioned as a one-dimensional "quick and dirty" approach to measure usability, research over the past 25 years revealed helpful insights and dimensions to contextualize and compare individual SUS scores on. In this paper, we present an open source web-based analysis toolkit for the SUS questionnaire, which calculates SUS measurements, analyses them based on the insights and contextualization scales suggested by previous work, and provides versatile plotting facilities to create appealing SUS graphs for scientific publications and presentations.

CCS CONCEPTS

• **Human-centered computing → Usability testing**.

KEYWORDS

Usability, Toolkit, SUS, System Usability Scale, Open Source, Free, Calculation, Contextualization, Plotting, Graphs, Figures

ACM Reference Format:
Jonas Blattgerste, Jan Behrends, and Thies Pfeiffer. 2022. A Web-Based Analysis Toolkit for the System Usability Scale. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3529190.3529216

1 INTRODUCTION

According to the technology acceptance model by Davis et al. [10], user form a behavioural intention to use products, prototypes, or software based on their attitude towards and consequent acceptance of them. This attitude is influenced by the perceived usefulness but also the perceived ease of use, also called *usability*. Consequently, usability is an important field of study where interdisciplinary perspectives and approaches meet. Besides many techniques, such as qualitative surveys, interviews, and observational techniques, one established method of measuring usability is using validated usability questionnaires. Here, a famous usability questionnaire is the System Usability Scale (SUS) by Brooke [7]. Originally envisioned and self-described as a one-dimensional "quick and dirty" approach, SUS questionnaires accounted for about 43% of post-study usability questionnaires used in the experiments identified in a meta analysis conducted by Lewis et al. [19] in 2009. Throughout the last 25 years, the initial validation of the questionnaire with $n = 20$ participants increased to $n = >10{,}000$, making the SUS a "fairly quick, but apparently not that dirty" approach as Lewis described it [18]. With recent developments towards the application of the SUS in new contexts such as elderly people or people with cognitive impairments [14] and the validation efforts of the SUS for various languages [12], this trend shows no sign of slowing down. Ultimately, the SUS has good reliability with a coefficient alpha usually around 0.92, high correlations with *likelihood to recommend* (0.75) and high correlations of *overall experience* (0.80) [4].

Besides its use in studies and specific validation endeavours of the SUS questionnaire itself, multiple researchers proposed approaches to contextualize SUS scores, trying to answer the question of what a specific SUS score actually means. As SUS scores, spanning between 0 and 100, follow neither a normal nor a uniform distribution, they cannot be interpreted linearly and especially not as a percentage value. Consequently, researchers calculated percentile curves of SUS scores from SUS study datasets, tried to contextualize SUS scores on adjectives, grading, net promoter score, quartile and acceptability scales, calculated at which point SUS scores become conclusive, and investigated the dimensionality of the SUS questionnaire by deriving learnability as a secondary dimension besides the usability of a system. All these contextualization and interpretation insights potentially add value over reporting pure SUS scores.

To date, only a handful of mostly commercial tools exist that help to calculate SUS scores. While these likely help usability researchers, the calculation of SUS scores itself is fairly simple and most tools do not provide further support, such as allowing researchers to compare different conditions, to plot graphs, to provide statistical analysis or to contextualize the calculated results regarding the aforementioned interpretation scales. Notably, comparatively sophisticated toolkits do exist for competing questionnaires, like the User Experience Questionnaire (UEQ) by Laugwitz et al. [16]. As we believe such features would be especially valuable for other usability researchers and practitioners using the SUS, we have developed an open source web-based analysis toolkit for the system usability scale that can calculate SUS scores, create a variety of different SUS plots and help researchers to contextualize their results on the interpretation scales developed in previous works. In contrast to the available tools, a special focus lays hereby on the usage of the tool in scientific studies and the report of results in

*Both authors contributed equally to this work, with Jonas Blattgerste predominantly contributing to the conceptual and Jan Behrends to the technical development.

PETRA '22, June 29-July 1, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3529216

the form of measurements and camera-ready figures in scientific publications. Additionally, we introduce the concept of normalizing the per-question averages of all 10 questions to represent their contribution towards the corresponding SUS study score to provide additional comparative insights.

2 RELATED WORK

While, to our best knowledge, holistic tools combining the calculation, comparison, contextualization and plotting of SUS scores do not exist to date and there is no scientific literature on such SUS calculation tools, there are some free and commercially available tools for the calculation of SUS scores with varying scopes (see Table 1). One free example is the "*SUS Calculator*" from *uiux-trends.com* [29], which allows for the calculation of an average SUS score based on an in-browser version of the SUS questionnaire. The free online appendix [2] of the book "*Measuring the user experience: collecting, analysing, and presenting usability metrics*" by Albert et al. [1] contains an Excel spreadsheet for the calculation of multiple SUS scores based on transcribed values from the SUS questionnaire using the SUS calculation formula.

A commercially available SUS tool by *quix.app* [21] can be included in websites and apps as a questionnaire and report average SUS scores on a dashboard with contextualization on grading and adjective SUS scales. Similarly, the commercial SUS tool included in the usability toolkit by *tryMyui.com* [27] allows for interpretation, contextualization and even rudimentary plotting of SUS scores based on online questionnaires used to investigate the usability of websites. The "SUS Guide & Calculator Package" provided as a commercially available Excel sheet for the calculation and interpretation of SUS scores by MeasuringU [20], a company founded by Jeff Sauro and James R. Lewis, allows for the SUS calculation, interpretation, statistical analysis and computation of sample size conclusiveness. While the commercially available "*System Usability Scale (SUS) Plus*" tool by *www.usabilitest.com* [30] has no functionality for plotting graphs, it is quite comprehensive in contextualizing SUS results based on transcribed values from the SUS questionnaire, including contextualization on the letter grading scale, adjective scale, net promoter scale, calculating Cronbach's alpha and the learnability dimension. It even allows for the generation of SUS questionnaire PDFs tailored to the website's or prototype's name. An analysis toolkit in the form of an Excel spreadsheet, provided supplementary to guidelines for usability research by the Victorian Government [13], allows for the calculation of average SUS scores for multiple conditions and their contextualization on the acceptability scale.

Not primarily envisioned as a calculation or plotting tool but rather aiding decisions about confidence interval problems based on the knowledge that the distribution of SUS study scores is generally skewed, which violates symmetry assumptions, Clark et al. [9] developed a freely accessible calculation tool providing recommendations for practitioners about which confidence intervals to apply to their results based on the sample size of the study. Additionally, the tool also calculates and plots the inserted SUS study contextualized on some SUS contextualization scales. Also, not primarily a SUS calculation tool but noteworthy, Xiong et al. [31] developed "SUSapp", a free iOS smartphone application, to aid the SUS usage for research or product development. Using their smartphone-based app, experimenters can hand the smartphone with SUS questionnaires for a product to participants, who fill out the questionnaire and hand the smartphone back to the experimenter, and then the app calculates the subject's individual SUS scores and the product's average SUS score.

3 THE SUS ANALYSIS TOOLKIT

Many usability practitioners using the SUS default to using tools like Excel to calculate SUS scores and plot corresponding graphs, which is not only cumbersome but could also lead to wrong interpretations, as raw SUS scores are easy to misinterpret, e.g., because they are not linear or to be interpreted as a percentage value. The tool proposed in this paper is supposed to support them by offering an effective and efficient way to calculate all relevant SUS measurements and create clearly visualized, interactable, and customizable graphs for provided SUS datasets (e.g. uploaded as CSV files). Practitioners and researchers can use the proposed tool to quickly evaluate their usability study results based on these contextualization approaches and insights, or generate and download publishable SUS graphs for research publications and presentations. In its current state, the analysis tool is capable of importing both single condition studies, but also the results of multiple SUS studies or conditions, e.g. for iterative formative studies, comparative studies with the SUS score as the independent variable or A/B testing.

Hereby, the tool offers several ways to configure and customize the generated interactive plots. Features among others include choosing between different plot styles (e.g. box plot, bar chart, radar charts, or percentile curves), showing only the results of specific conditions or questions, choosing between different contextualization scales, showing a conclusiveness graph based on the number of participants for a condition, or showing average scores for specific SUS items. All of these generated plots are downloadable as graphs in PNG formats.

3.1 User Workflow

The SUS Analysis Toolkit is purely web-based and no installation or additional software is necessary, the tool can be directly accessed in any browser. Notably, it can also be hosted locally through a Python server using the open source code of the toolkit, if desired. To import the SUS scores of a usability study, they have to be converted into a CSV file. In this file, one row represents a filled-out questionnaire, with each column being one of the ten raw item values of the SUS and finally the eleventh row being a variable for the condition or study. The starting page of the tool provides examples of valid CSV files for both multi- and single variable usability evaluations, which can be downloaded and used as a template. Furthermore, the starting page contains a short introduction and explanation of the tool, as well as relevant references to the scientific publications the tool is based on. When the CSV file has been created, it can be uploaded directly to the starting page by drag-and-dropping the CSV file into one of two upload forms, one for a single variable usability evaluation and the other for the multi-variable evaluations. After the CSV file is validated, e.g. checked for the correct format and plausibility of raw scores, the user will either be redirected to the multi- or single study interface, depending on which upload

Tool	Licensing	Calculation	Comparison	Benchmarking	Plotting	Validation	Statistics
uiuxtrends.com [29]	freeware	(✓)					
Albert et al. [2]	freeware	(✓)					
quix.app [21]	commercial	✓		✓			
tryMyui.com [27]	commercial	✓		✓			
measuringu.com [20]	commercial	✓	✓	✓	(✓)	✓	✓
usabilitest.com [30]	commercial	✓		✓		(✓)	
vic.gov.au [13]	freeware	✓	✓	(✓)			
Clark et al. [9]	open source	✓		✓	(✓)		(✓)
Xiong et al. [31]	open source	✓					(✓)
SUS Analysis Toolkit	open source	✓	✓	✓	✓	(✓)	(✓)

Table 1: The 9 existing SUS tools identified during our investigation and the SUS Analysis Toolkit itself, visualized with their licences and provided functionality in terms of calculation, comparison, contextualization, plotting, validation and statistics.

form was chosen. This process is visualized in Figure 1. The multi-variable interface will be described in more detail in Section 3.2 and the single variable interface in Section 3.3.

Depending on the size of the conducted study, the most time-consuming part of this procedure is likely the creation of the CSV file by transcribing the raw SUS values from the questionnaire, which might take some time. Once this is done and uploaded to the tool, the creation of SUS graphs themselves is instant and the customization of graphs should not take more than a few minutes. All generated plots and tables can be downloaded as camera-ready PNG graphs and CSV tables independently, customized to fit their intended application case (e.g. single- or double-column papers), or downloaded as a complete analysis.

3.2 Multi-Variable Interface

If the user chose the multi-variable upload option, he is consequently redirected to the multi-variable interface shown in Figure 2. The multi-variable upload is supposed to offer a customizable utility toolkit to compare the results of multiple SUS studies or conditions within comparative evaluation studies using the SUS score as an independent variable with each other. Here, the users can retrieve all relevant quantitative measurements (e.g. average, median, quartiles, min, and max values) for each of the variables, view different analyses (e.g. comparing conditions, comparing individual questions, or checking the conclusiveness) and then customize all displayed plots according to their need and download them as graphs and tables. The multi-variable SUS analysis interface consists of the following plot/table combinations with their respective customization options, which are explained in the following subsections: SUS study score comparison, SUS study score percentile curve contextualization, Per-Item/Question score contributions, and conclusiveness of the results.

3.2.1 SUS Study Score Comparison. The SUS Study Score Comparison Graph is the closest to traditional SUS graphs found in existing literature, visualizing SUS study scores as either box plots, notched box plots, or bar charts and allowing the contextualization of scores using contextualization scales (see Figure 3 and Figure 4 for the corresponding data table). Through the customization options on the right side of the interactive plot (see Figure 2 - Customization Options), users may customize their graphs. Customization features include, but are not limited to:

- the plot type (barchart, box plot, or notched box plot) of the graph
- comparison scales: Adjective scale, grade scale, quartile scale, acceptability scale, net promoter Scale
- which of the different variables in the uploaded CSV-file should be plotted
- whether all data points, only outliers or neither should be plotted
- whether the mean, standard deviation, both or neither should be plotted
- the orientation (horizontal or vertical) of the graph
- the title of the X-Axis of the graph (E.g. describing the difference between the variable)

The customized interactive plot can be utilized by users to compare the average SUS score of multiple conditions in a single study or iterative SUS study scores for formative testing purposes with each other and with several provided types of contextualization scales. The contextualization scales, their origins, and scientific derivations are as follows.

Adjective Scale: The adjective scale (see Figure 3 on the right side of the plot) consists of the adjective ratings „Worst Imaginable", „Poor", „OK", „Good", „Excellent" and „Best Imaginable", where each of the 1-100 SUS scores corresponds to one of these adjectives. The measurements for this scale were taken from Sauro [24], which in turn is based on a study by Bangor et al. [3] in which a large (n=1000) usability study was conducted where a seven-point likert scale with the adjectives was added to the SUS questionnaire. The results of the study showed a high correlation between six of the adjectives of the likert scale and the overall SUS scores, subsequently creating the adjective scale.

Grade Scale: Bangor et al. [3] suggested using the traditional school grading scale (i.e., 0-59 = grade F, 60-69 = D, 70-79 = C, etc.) to contextualize SUS scores. This seems to be a natural fit, since according to the study, a SUS score of 70 is about average and 70 on the school grading scale would correspond to a C -, generally perceived as an average grade. However, this method has limitations. For example, it is almost impossible to get an A,

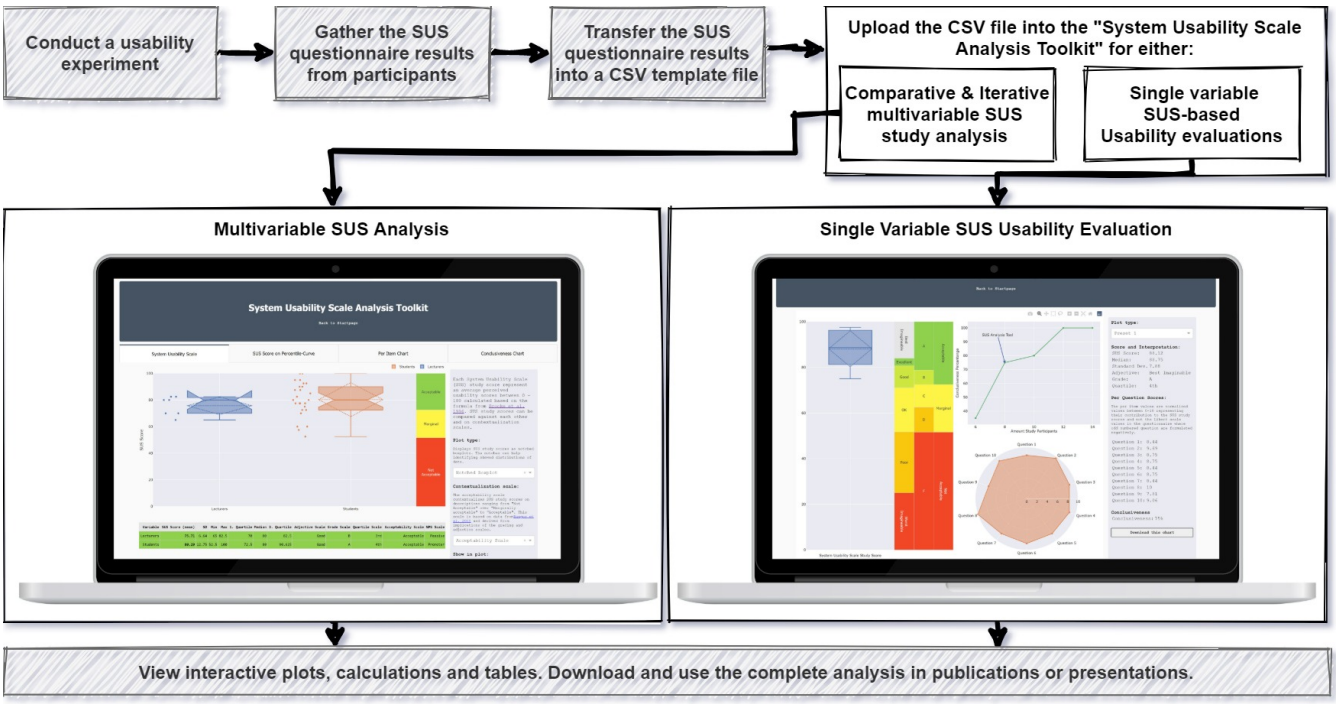


Figure 1: The user flow of the SUS Analysis Toolkit: After conducting an experiment utilizing the SUS questionnaire, results can be transferred into a CSV template file provided by the tool. Depending on the type of study (e.g. an iterative or comparative usability study with SUS scores being the independent variable or a singular usability evaluation utilizing the SUS questionnaire), the CSV file can be uploaded to the single- or multivariable SUS Analysis section. After calculating all relevant metrics and generating customizable, interactive plots, the user can download and use the generated calculations, tables and graphs.

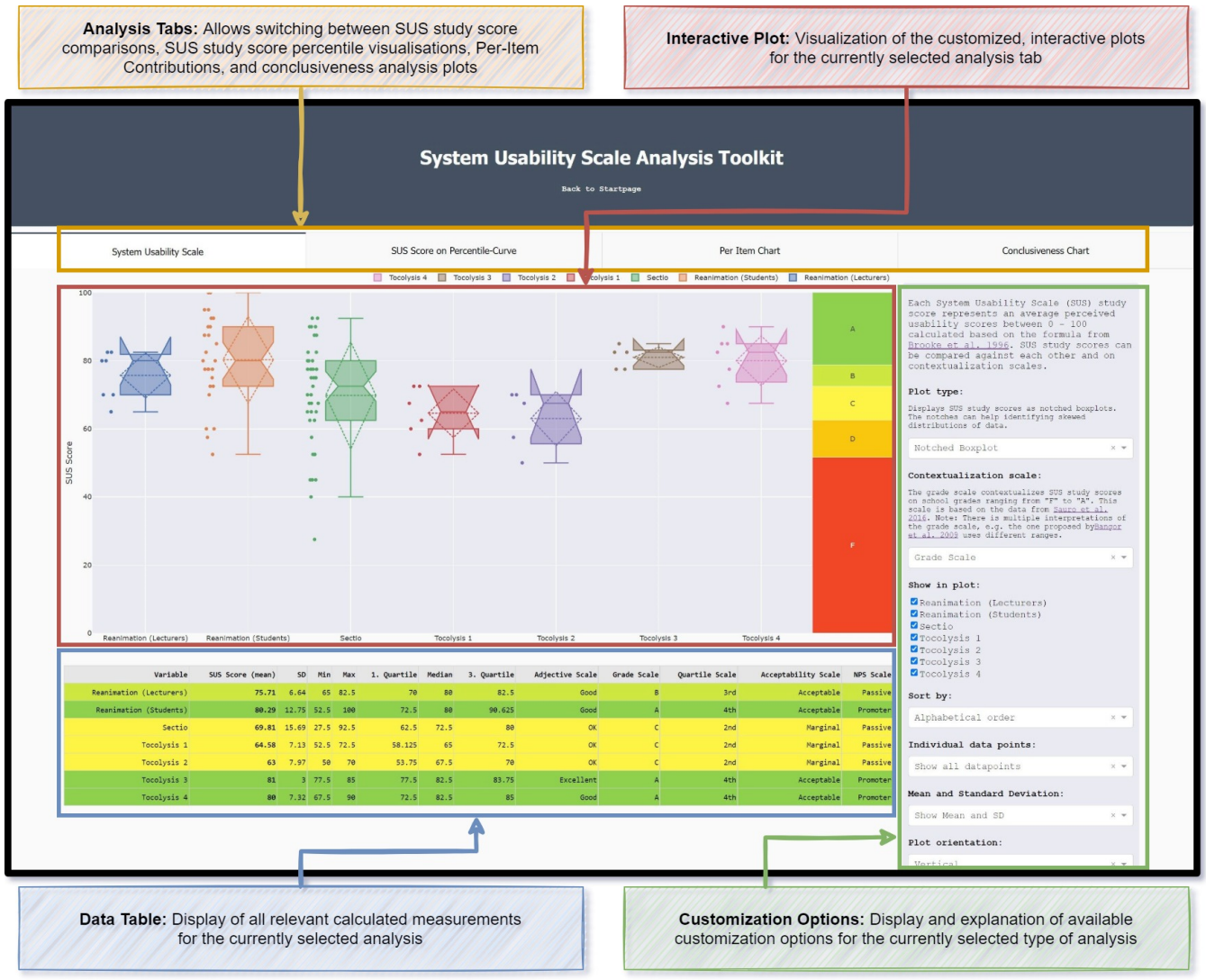
since SUS scores above 90 are exceedingly rare, with only 1% of SUS scores [22] reaching them as can be seen on the percentile curve graph displayed in Figure 5. Therefore, this original grading scale was later revised by Sauro et al. [25]. Here, they analysed data from 446 surveys and 5000 individual SUS responses to create a revised grading scale based on the percentiles of the data. This revised grading scale is used in the tool. (see Figure 2 on the right side of the interactive plot area)

Quartile Scale: The quartile scale contextualizes the SUS score on the quartile scale identified by Bangor et al. [3] that was also used to develop the adjective and grading scale, visualizing all 4 quartiles and the median of the dataset. Importantly, this is to be distinguished from the percentile curve and the corresponding percentile curve graph, which was developed based on a dataset and analysis from Sauro et al. [25].

Net Promoter Scale: Another metric that has been shown to correlate with the SUS is the Net Promoter Score (NPS). The NPS is a widely used, yet arguably controversial, metric of customer loyalty. It consists of a single question: "How likely is it that you would recommend our company to a friend or colleague?". This question is then answered on an eleven-point likert scale ranging from "Not likely at all" to "Extremely likely". Responders who give responses from 0 through 6 are called "Detractors" and those who give responses at 9 or 10 are called "Promoters", the remaining are called "Passives". Sauro [23] shows, that there is a correlation between the SUS and the NPS and consequently created the NPS scale.

Acceptability Scale: Finally, the acceptability scale classifies average SUS scores as either „acceptable" or „not acceptable". According to Bangor et al. [4], as the average SUS Score is roughly 70, SUS scores higher than 70 would be „acceptable", while those below 50 are „not acceptable". The range between 50-70 is „marginally acceptable", divided into „low marginal" and „high marginal". The lower ranges for this scale were chosen by considering the ranges of the grading and adjective scale. For example, on the adjective scale, a score of 51.6 is „OK", so anything lower is considered „not OK" and subsequently „not acceptable". (see Figure 8, the third contextualization scale from the left)

3.2.2 Percentile Curve Contextualization Graph. This graph shows the calculated SUS study scores of the uploaded dataset on a percentile curve derived from over 5000 SUS questionnaires. The data for this curve is taken from Sauro et al. [25]. Viewing the results of the uploaded SUS study on this curve is not redundant to the traditional SUS comparison graph, not even when using the quartile contextualization scale. Importantly, the quartile contextualization scale is based on a different dataset by Bangor et al. [3], which can lead to small differences when comparing the SUS study score to percentile correlations between the two calculation approaches. Using the percentile curve tells the researcher how well the systems



Variable	SUS Score (mean)	SD	Min	Max	1. Quartile	Median	3. Quartile	Adjective Scale	Grade Scale	Quartile Scale	Acceptability Scale	NPS Scale
Reanimation (Lecturers)	75.71	6.64	65	82.5	70	80	82.5	Good	B	3rd	Acceptable	Passive
Reanimation (Students)	80.29	12.75	52.5	100	72.5	80	90.625	Good	A	4th	Acceptable	Promoter
Sectio	69.81	15.69	27.5	92.5	62.5	72.5	80	OK	C	2nd	Marginal	Passive
Tocolysis 1	64.58	7.13	52.5	72.5	58.125	65	72.5	OK	C	2nd	Marginal	Passive
Tocolysis 2	63	7.97	50	70	53.75	67.5	70	OK	C	2nd	Marginal	Passive
Tocolysis 3	81	3	77.5	85	77.5	82.5	83.75	Excellent	A	4th	Acceptable	Promoter
Tocolysis 4	80	7.32	67.5	90	72.5	82.5	85	Good	A	4th	Acceptable	Promoter

Figure 2: The four basic components of the System Usability Scale Analysis Toolkit User Interface: The Analysis Tabs, Interactive Plots, Data Tables, and Customization Options. In this case, the SUS Score study comparison analysis tab is selected and showing the interactive notched box plot and corresponding data table according to the chosen customization options. The displayed data is a real example from usability evaluations conducted during project Heb@AR [5, 6].

compare to each other contextualized with systems in the dataset from Sauro et al. [25]. This makes the non-linear nature of the SUS clearly visible, where small increases in average SUS scores can actually make comparatively large differences when contextualizing the scores with those of other studies and subsequently imply observable differences in usability. See Figure 5 for an example of the percentile curve graph generated with the SUS Analysis Toolkit.

3.2.3 Per-Item Contribution Graph. While the SUS comparison graph shows comparisons between the average SUS study scores of the variables, the Per-Item Contribution Graph allows for individual question comparisons of the SUS. It shows the average contribution of each of the 10 questions of the SUS questionnaire towards the average SUS study scores. (see Figure 6) This means, while the questionnaire questions are alternating between positive/negative statements, the graph is hereby already normalized, so higher scores correspond to better results and more contribution towards the SUS study score. The idea behind this graph is for the researcher to be able to evaluate and compare individual aspects of the questionnaire, by only regarding the items corresponding to the area of interest or visualizing significant differences in specific perceptions of the user. For example, there is previous work suggesting that the SUS, in addition to usability, could also measure learnability by only analysing two out of the ten questions, items 4 and 10.

Nonetheless, practitioners should be careful with such evaluations, since the SUS in general is only meant to be interpreted as an

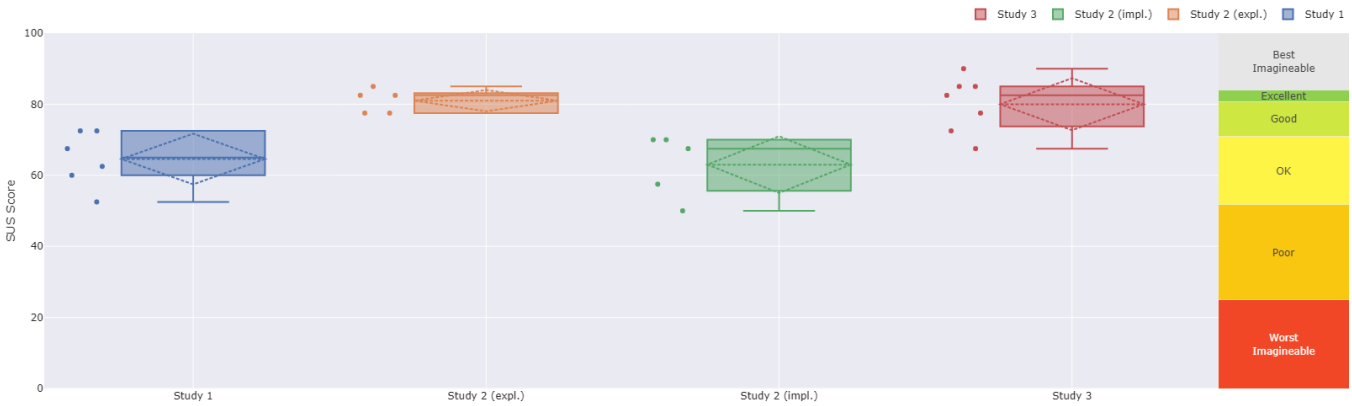


Figure 3: An example of the SUS score comparison plot with 4 independent variables from an iterative usability study. This data is a real example from Blattgerste et al. [6]. In this case, the plot is customized to display a box plot with mean, standard deviation, and individual data points. On the right side, the "adjective" scale from [4] is displayed for contextualization.

Variable	SUS Score (mean)	SD	Min	Max	1. Quartile	Median	3. Quartile	Adjective Scale	Grade Scale	Quartile Scale	Acceptability Scale	NPS Scale
Study 1	**64.58**	7.13	52.5	72.5	58.125	65	72.5	OK	C	2nd	Marginal	Passive
Study 2 (expl.)	**81**	3	77.5	85	77.5	82.5	83.75	Excellent	A	4th	Acceptable	Promoter
Study 2 (impl.)	**63**	7.97	50	70	53.75	67.5	70	OK	C	2nd	Marginal	Passive
Study 3	**80**	7.32	67.5	90	72.5	82.5	85	Good	A	4th	Acceptable	Promoter

Figure 4: Besides the comparative plot, the most important calculated metrics are also displayed in form of a table. In this case, the table for the plot shown in Figure 3, it displays all contextualization scale results and additionally colours the results according to the currently selected scale for the plot: the "adjective" scale.

overall score [7], individual questions are not supposed to provide diagnostic value in themselves, and they do not relate to specific features of evaluated systems [8]. Customization features for the Per-Item Contribution Graph include, but are not limited to:

- the plot type (barcharts, radar chart or stacked bar chart) of the graph
- which of the different variables in the uploaded .csv-file should be plotted
- which individual items should be plotted
- the orientation (horizontal or vertical) of the graph

3.2.4 Conclusiveness Graph. The conclusiveness graph (see Figure 7) is based on a study conducted by Tully et al. [28], in which they tried to determine, which sample sizes were needed for different usability questionnaires to be conclusive, one of these being the SUS. To do this, they took varying amounts of completed SUS questionnaires (6, 8, 10, 12, 14) and observed at what point the average of these scores came to the „correct" conclusion, with the correct conclusion here being the average of the full data set. They found that with a sample size of 8, a SUS study is already 75%, with 10 80% and with 12 or more 100% conclusive. The conclusiveness graph in the SUS Analysis Toolkit takes the number of participants per condition of the uploaded CSV file and plots them on a graph derived from the conclusiveness percentages. Due to the unavailable data in Tully et al. [28] for sample sizes smaller than 6 participants, no conclusiveness for such sample sizes can be calculated and are displayed as 0% in the tool.

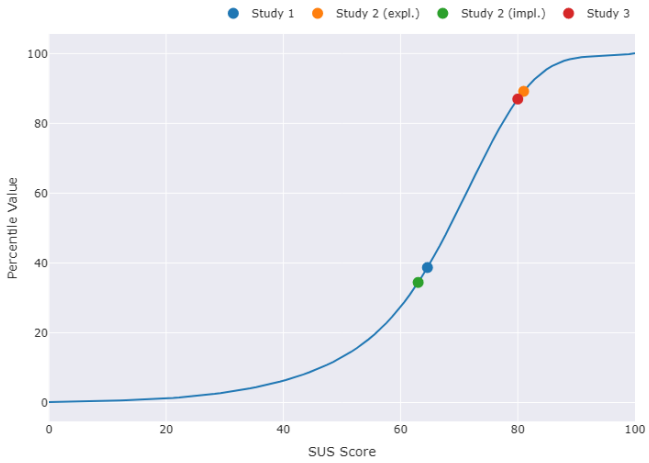


Figure 5: The percentile curve graph, plotted by the SUS Analysis Toolkit based on Sauro et al. [25] with data taken from Blattgerste et al. [6].

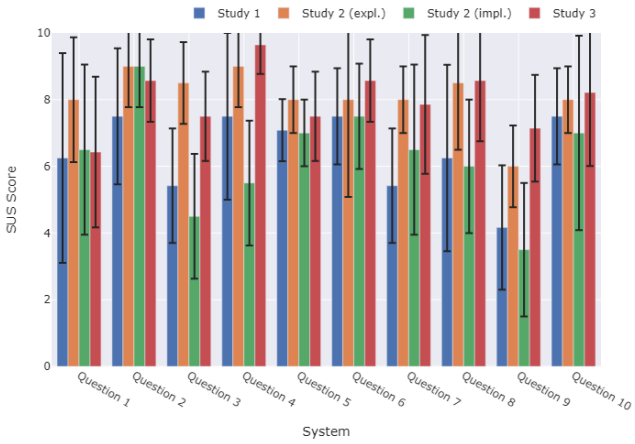


Figure 6: The per-item contribution graph, plotted by the SUS Analysis Toolkit with data taken from Blattgerste et al. [6]. This graph shows the contribution of each individual question towards the SUS study scores.

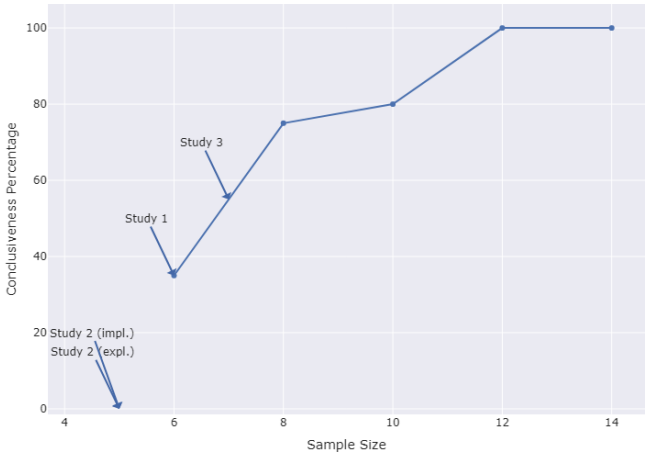


Figure 7: An example of the conclusiveness graph plotted by the SUS Analysis Toolkit based on Tullis et al. [28] with data taken from Blattgerste et al. [6]. In this case, no conclusiveness percentage can be calculated because of the small sample size of 5 participants for 2 of the variables.

3.3 Single Variable SUS Dashboard

The single variable SUS dashboard interface gives a quick and broad overview over the perceived usability results from a study where the SUS score was the dependent variable without another condition for between or within subject comparisons using the SUS study scores. An example of the Single Variable SUS dashboard is shown in Figure 8, that displays the SUS results from the formative usability study of the SUS Analysis Toolkit itself. Through the contextualization scales, conclusiveness, percentile ranking, and per-item contribution graph for the SUS study score, there can still insights be derived from this singular SUS study score. To achieve this, multiple plots with different emphasis are displayed in one dashboard-style figure, as well as the most relevant quantitative data in a table. While no direct customization options are given to the user, they may choose between different presets for these dashboards. The single variable dashboard graph consists of the same plots that are used in the multi-variable interface, but since it is meant to produce a quick and dirty dashboard-like graph consisting of the most important information and insights that can still be derived from a singular SUS study score, it does not have the same focus on customization options. The user may choose from the four available presets, get the numerical results of the calculations, and download the generated graphs as PNG files:

- SUS study score + Per-Item Radar Chart + Conclusiveness
- SUS study score + Per-Item Likert Scale + Conclusiveness
- SUS study score + Per-Item Likert Scale + Percentile Curve
- SUS study score + Percentile Curve + Conclusiveness

4 EVALUATION

To gather first impressions and feedback for the toolkit from the actual target group of the toolkit during the development process, a formative usability study was conducted with 8 usability practitioners and researchers that were aged 25 to 56 years, with an average age of 34 years (SD = 9.64). One participant was female. After acquiring demographic data, participants were presented with a fictional scenario incorporating specific tasks, designed to be solvable using the SUS Analysis Toolkit and covering most of its functionality. For example, in one of the tasks, participants had to calculate the SUS study scores and then find specific quantitative parameters of the conditions to conclude which of the conditions performed the best. For this, participants were provided a prepared CSV file, which they had to upload into the tool. (see Figure 1) While the participants performed the provided tasks, they were instructed to verbalize their thought process in line with the think-aloud methodology and observed via screen sharing. After participants were done with all tasks, they were asked to fill out a short qualitative questionnaire, the SUS questionnaire, and an S-UEQ questionnaire [26]. Overall, this formative usability study showed that there is an interest in the tool: Multiple participants asked to be kept up-to-date regarding the development status of the SUS analysis toolkit. The results of the SUS (Score: 88.13, SD=8.83, and Conclusiveness: 75% as visualized in Figure 8) and S-UEQ (hedonic score=1.28, pragmatic=2.00) questionnaires suggest, that the usability and user experience of the tool were perceived to be excellent. The verbal think-aloud feedback in particular helped to generate multiple specific improvements regarding the user interface. Most feedback was with regard to smaller inconveniences like the wish for more consistent numbering of the questions in the Per-Item graph or the missing of convenience options like sorting across the graph customizations, which were both included into the now released version of the tool. Furthermore, the qualitative questionnaire showed a wish for more functionality regarding statistical analysis, which was also brought up in the verbal feedback. This formative evaluation was conducted on a previous version and the SUS Analysis Toolkit, which has been updated based on the provided feedback but also additional non-representative feedback sessions since.

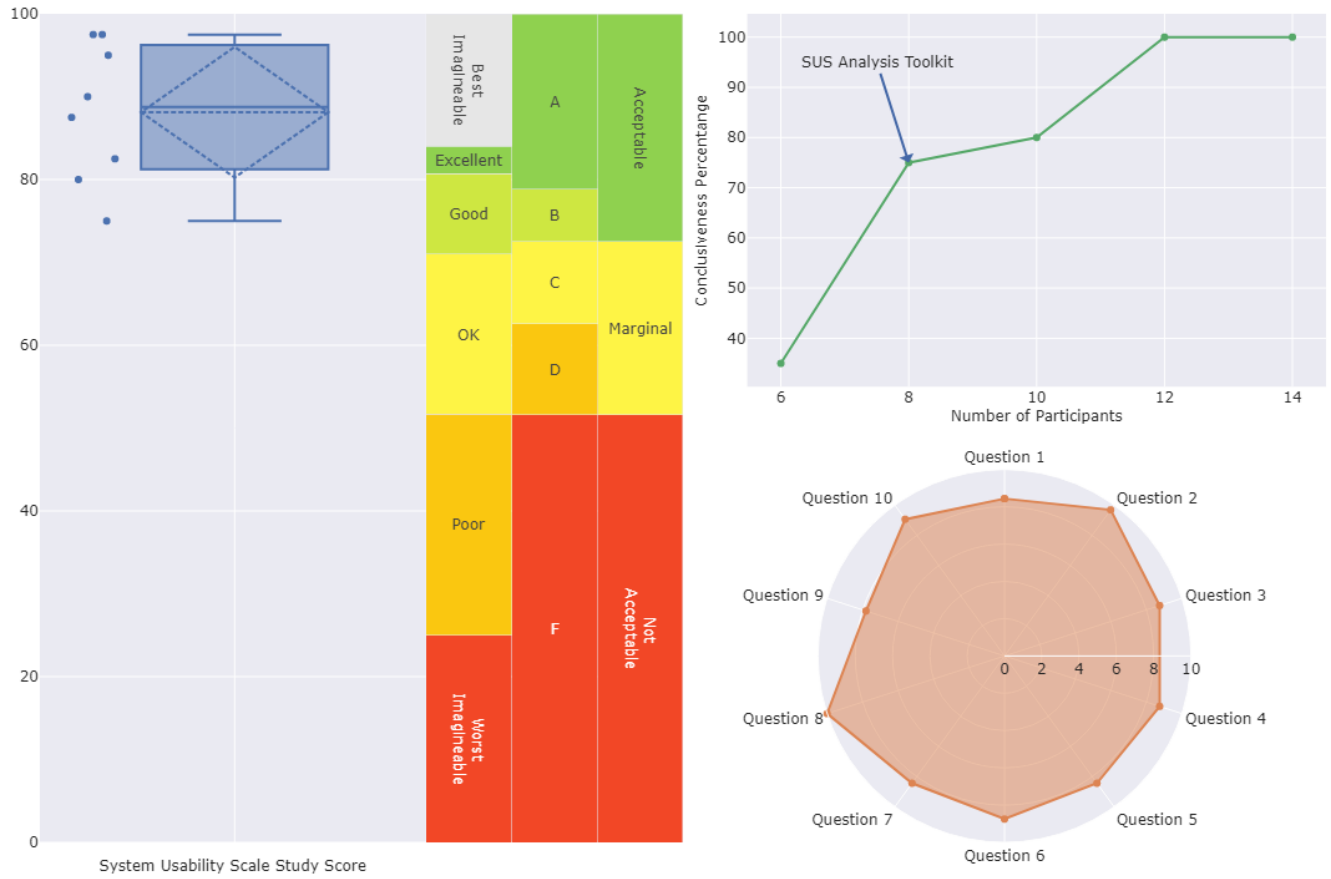


Figure 8: An example of the single variable dashboard graphs, plotted by the SUS Analysis Toolkit. This one displays the SUS study score of the formative evaluation of the SUS Analysis Toolkit itself. On the left side, a box plot with mean and standard deviation is displayed next to the raw data points and three of the contextualization scales, on the top right, the datasets' conclusiveness is shown and on the lower right the individual questions' contribution towards the SUS study score is visualized.

4.1 Summative Evaluation

As the formative evaluations performed during the development process already resulted in excellent usability scores and qualitative feedback indicating this preliminary version would already have satisfied most of the users needs, we believe a summative evaluation of the improvements made under lab conditions would currently not further add additional insights. Rather, we would like to utilize the open source nature of the project, publish the tool, this corresponding publication explaining the ideas and science behind it, and the repository containing the open source code. We will then start an evaluation period in line with the field usability testing methodology [11], where other researches can deploy the tool and provide feedback (e.g. through email or GitHub Issues) from their actual usage experience for further improvements or future directions of the toolkit.

5 DISCUSSION

Regarding this toolkit, we were our own first customer. We initially created the tool because we needed a versatile calculation and plotting utility for, in our case, the usability evaluation of pervasive technologies and immersive interaction concepts, a field where determining what implementations lead to certain usability outcomes is generally not well known yet, making both comparative usability studies but also iterative usability comparisons of utmost importance for the success of applications and concepts. The SUS with its seniority and simplicity seemed to be a good fit, and previous work provided contextualization approaches and insights that would add extra value. Nonetheless, a tool effectively and efficiently combining those in a way that would fit our needs was simply missing.

That the SUS Analysis Toolkit does not only fill this gap and fits our own needs but can improve the workflow for all researchers and practitioners using the SUS in usability studies is clear through the formative evaluation study feedback of the tool but also the application of it by partners close to our research group. As we

want to add value for even more researchers and practitioners, we now make it available, open source and freely accessible.

5.1 Novel Contributions

While the main contribution of this work and the SUS Analysis Toolkit is the gathering and combination of existing insights, contextualization approaches, and analysis from previous work into a single toolkit that exceeds existing approaches and making it accessible for free, there are more novel contributions to this work than it just being the sum of existing parts.

To our best knowledge, the analysis and plotting of individual items of the questionnaire through normalizing them to represent their contribution towards the corresponding SUS study score has not been explored before. While Brook [8] warned that individual questions have no inherent meaning and are not diagnostic, this does add value. While it is true that the individual questions are not inherently diagnostic, having the per-item scores in the form of their contribution towards the SUS study scores (10 questions with a possible contribution ranging from 0 to 10) allows for comparisons, e.g. on bar charts, radar charts or stacked bar charts that can provide valuable insights in iterative or comparative usability studies. These differences would not be visible or could be easily confusing when the original likert scale items, which alternate between positive and negative statements, are plotted or used for comparative purposes.

Additionally, the single-variable dashboard plot, allowing for efficient creation and instant contextualization of usability evaluations to available contextualization scales in combination with also displaying percentile ranking, conclusiveness and/or per-item contributions in a single graph by simply selecting a pre-assembled preset is a novel contribution, extending previous ideas of combining several contextualization scales into a single graph, started by Bangor et al. [3], towards combining all relevant insights holistically.

5.2 The Learnability Dimension

The SUS Analysis Toolkit incorporates almost all contextualization insights available through previous research, beside one exception. It was briefly considered to have the tool not only evaluate the overall usability of a system, but to look at individual items of the SUS and attempt to evaluate more specific parts of the system. For example, two items of the SUS ask about the learnability of a system (Item 4: „I think that I would need the support of a technical person to be able to use this system.“, Item 10 : „I needed to learn a lot of things before I could get going with this system“). Even though the SUS' creator John Brooke warned that individual items of the SUS are not meaningful on their own [8], Lewis et al. [19] explored, whether it would be possible to derive a learnability dimension from the SUS in 2009 by only interpreting questions 4 and 10 together. In a later paper from 2017, however, Lewis et al. [17] take a more detailed look at their own proposition in the form of an exhaustive meta-analysis, after which they conclude that the SUS is after all not a significant measure of learnability. Therefore, this idea was disregarded, and the tool only evaluates for usability.

5.3 Open Source Availability

The tool is currently hosted on our server and accessible through the URL https://sus.mixality.de and can be used freely for any purpose. Additionally, the complete source code of the SUS analysis toolkit is published at https://github.com/jblattgerste/sus-analysis-toolkit under the MIT Licence as an open source project. For one, this allows other researchers to validate our calculations and interpretations, but also allows for a lasting impact of the tool through contributions from third parties that continuously improve it.

5.4 Ownership of generated content

As both, the SUS Analysis Toolkit proposed in this paper, but also the underlying Dash and Plotly libraries [15] are free, open source, and licensed under the MIT licence, no restrictions or automatic licensing applies for content generated with the tool. The ownership of created calculations, interpretations, tables, and plots fully remain with the user of the tool. Created content can be used for any purpose, including commercial use, without restrictions and with or without attribution to the SUS analysis toolkit.

5.5 Privacy Considerations

For many usability researchers and practitioners, privacy considerations are of importance. While data imported into the tool is inherently at least anonymized based on the accepted CSV data structure consisting of the values for the 10 questions of the SUS combined with an independent variable name, uploading study data onto a server might not be appreciated or not possible based on the gathered participation consent of studies. This might even remain problematic if the source code of the tool is fully accessible.

Therefore, a design was chosen where data is shared with a server but never stored on it. The data imported into the tool is uploaded to the server for the plotting of the figures, but is only temporarily stored in a local data frame in the users' browser. The tool utilizes this local data frame for its analysis by uploading the data and chosen options to the server. The server applies the calculations and selected plotting options to the data and returns the interactive plot to the users' browser. The data sent to the server is erased with the delivery of the plot. All data temporarily stored in the users' browser, besides downloaded content, is automatically erased when closing the browser tabs of the tool.

If this is not sufficient, the tool can also be run on a local server on a desktop PC. The admittedly slightly more complex process is described in the quick-start guide of the readme file provided with the open source code of the toolkit.

5.6 Generating SUS PDF questionnaires

There are more than two dozen versions of the SUS questionnaire that exist today. This makes it callenging to keep track of their origin, validity, reliability and compatibility with the toolkit. Therefore, as a side contribution, the SUS Analysis Toolkit also includes an open source SUS PDF generator, published under the MIT licence: https://github.com/jblattgerste/sus-pdf-generator. It can be accessed at https://jblattgerste.github.io/sus-pdf-generator/ and can be used to generate interactive SUS PDF questionnaires for more than 18 different versions, variations and languages, including customization options to e.g. use product names as the variable.

6 CONCLUSION

The SUS Analysis Toolkit in its current state is fully functional, hosted on a server of the Mixed Reality Research Group at University of Applied Sciences Emden/Leer, accessible for free through the domain https://sus.mixality.de, and fully open sourced at https://github.com/jblattgerste/sus-analysis-toolkit. It can be used to calculate, analyse, interpret, contextualize, and plot SUS scores from either singular, comparative, or iterative usability studies based on insights gained in previous SUS literature. In a formative usability study, participants signalized interest in using the toolkit in their own work and the toolkit is already used in several publications and in forthcoming and ongoing usability evaluations inside and close to our research group.

6.1 Limitations & Future Work

While the scales and insights utilized to contextualize the calculated SUS scores are the most helpful feature of the tool and are validated through sufficiently big datasets, they are also its biggest limitation. Only if the underlying data and contextualization interpretations from previous work are correct, the tool provides useful, correct answers. As discussed in Section 5, e.g. for learnability, the potential second dimension of the SUS questionnaire, proposed dimensions, insights, and contextualization scales not always result in significant differences when analysed after extensive utilization.

Besides this limitation and ongoing feedback and suggestions by users of the tool after its deployment, we plan to expand export options and formats and develop an API which can be used to import data into the tool remotely and continuously update scores and conditions, resulting in live plotting and contextualization of graphs and tables. Additionally, the current scope of the tool is the usage as a web-based toolkit in a browser of a desktop PC. While technically already possible, we plan to improve the user interface towards additionally incorporating mobile platforms like tablets and smartphones. Finally, as mentioned by some participants in the evaluation, while some descriptive statistics are calculated, inferential statistical analyses of the results are currently missing in the toolkit. We plan to explore this option with domain experts. Reviewing available tools and the recent literature, there are promising statistical insights which could add additional benefits. Beside statistical analyses of significant differences of variables, for example, the "SUS Guide & Calculator Package" by MeasuringU [20] provides functionally to statically compare SUS scores, check for inconsistent respondents, check internal reliability and provides sanity checks to prevent miscoding. In line with this, the decision rules introduced by Clark et al. [9] to bolster confidence interval accuracy, could also provide benefits.

ACKNOWLEDGMENTS

This research was partially supported by the grant 16DHB3021, project "HebAR - AR-Based-Training-Technology", by the German Ministry for Education and Research (BMBF).

REFERENCES

[1] William Albert and Thomas Tullis. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics.* Newnes.
[2] William Albert and Thomas Tullis. 2019. Calculating a System Usability Scale (SUS) score spreadsheet. http://measuringux.com/. Accessed: 2022-01-12.
[3] Aaron Bangor, Phil Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Stud.* 4 (04 2009), 114–123.
[4] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.
[5] Jonas Blattgerste, Kristina Luksch, Carmen Lewa, Martina Kunzendorf, Nicola H Bauer, Annette Bernloehr, Matthias Joswig, Thorsten Schäfer, and Thies Pfeiffer. 2020. Project Heb@ AR: Exploring handheld Augmented Reality training to supplement academic midwifery education. *DELFI 2020–Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik eV.*
[6] Jonas Blattgerste, Kristina Luksch, Carmen Lewa, and Thies Pfeiffer. 2021. TrainAR: A Scalable Interaction Concept and Didactic Framework for Procedural Trainings Using Handheld Augmented Reality. *Multimodal Technologies and Interaction* 5, 7 (2021), 30.
[7] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189 (1996).
[8] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.
[9] Nicholas Clark, Matthew Dabkowski, Patrick J Driscoll, Dereck Kennedy, Ian Kloo, and Heidy Shi. 2021. Empirical Decision Rules for Improving the Uncertainty Reporting of Small Sample System Usability Scale Scores. *International Journal of Human–Computer Interaction* (2021), 1–16.
[10] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.
[11] Joseph S Dumas, Joseph S Dumas, and Janice Redish. 1999. *A practical guide to usability testing.* Intellect books.
[12] Meiyuzi Gao, Philip Kortum, and Frederick L Oswald. 2020. Multi-language toolkit for the system usability scale. *International Journal of Human–Computer Interaction* 36, 20 (2020), 1883–1901.
[13] The Victorian Government. 2021. Analysing your findings: Find out how to turn your raw data into meaningful insights. https://www.vic.gov.au/stage-3-analysing-your-findings. Accessed: 2022-01-12.
[14] Richard J Holden. 2020. A simplified system usability scale (SUS) for cognitively impaired and older adults. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, Vol. 9. SAGE Publications Sage CA: Los Angeles, CA, 180–182.
[15] Plotly Technologies Inc. 2015. *Collaborative data science.* Montreal, QC. https://plot.ly
[16] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group.* Springer, 63–76.
[17] James Lewis and Jeff Sauro. 2017. Revisiting the Factor Structure of the System Usability Scale. *Journal of Usability Studies* 12 (08 2017), 183–192.
[18] James R Lewis. 2020. Perceived Usability - Usefulness and measurement. https://www.torchi.org/resources/Documents/2020-Apr-7%20Jim%20Lewis%20-%20Perceived%20Usability%20-%20slides.pdf. Accessed: 2022-01-12.
[19] James R Lewis and Jeff Sauro. 2009. The factor structure of the system usability scale. In *International conference on human centered design.* Springer, 94–103.
[20] MeasuringU. n.d.. SUS Guide & Calculator Package. https://measuringu.com/product/suspack/. Accessed: 2022-01-12.
[21] Quix. 2021. Track your usability & satisfaction with the System Usability Scale. https://www.quix.app/system-usability-scale. Accessed: 2022-01-12.
[22] J. Sauro. 2011. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices.* CreateSpace Independent Publishing Platform.
[23] Jeff Sauro. 2012. *Predicting Net Promoter Scores from System Usability Scale Scores.* https://measuringu.com/nps-sus/
[24] Jeff Sauro and Jim Lewis. 2018. *5 Ways to Interpret a SUS Score.* https://measuringu.com/interpret-sus-score/
[25] Jeff Sauro and James R. Lewis. 2012. *Quantifying the User Experience: Practical Statistics for User Research* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 203–204 pages.
[26] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (01 2017), 103. https://doi.org/10.9781/ijimai.2017.09.001
[27] TryMyUI. n.d.. SUS: The System Usability Scale. https://www.trymyui.com/sus-system-usability-scale. Accessed: 2022-01-12.
[28] Thomas Tullis and Jacqueline Stetson. 2006. A Comparison of Questionnaires for Assessing Website Usability. (06 2006).
[29] UIUXTrend. n.d.. SUS Calculator. https://uiuxtrend.com/sus-calculator/. Accessed: 2022-01-12.
[30] Usabilitest. n.d.. System Usability Scale (SUS) Plus. https://www.usabilitest.com/system-usability-scale. Accessed: 2022-01-12.
[31] Jeffrey Xiong, Claudia Ziegler Acemyan, and Philip Kortum. 2020. SUSapp: A Free Mobile Application That Makes the System Usability Scale (SUS) Easier to Administer. *Journal of Usability Studies* 15, 3 (2020).