A Metric to Assess the Reliability of Crowd-sourced SUS Scores: A Case Study on the PoPLar Authentication Tool

YONAS LEGUESSE, University of Malta, Malta
MARK VELLA, University of Malta, Malta
CHRISTIAN COLOMBO, University of Malta, Malta
JULIO HERNANDEZ-CASTRO, Technical University of Madrid, Spain

The concern of inattentive respondents in surveys is widely acknowledged and has been extensively researched, and crowd-sourcing platforms further complicate this issue with the additional problem of bot usage to automatically respond to surveys. This work explores this issue within the critical domain of usable security, highlighting limitations of the use of crowd-sourcing platforms for usability studies, particularly in the context of obtaining valid and reliable System Usability Scale (SUS) scores.

While crowd-sourcing platforms may already offer built-in quality controls, for example ensuring a respondent's positive historical performance with regards to the completion of previous surveys, our exploratory surveys showed that issues with bots and careless respondents persist. Building upon these insights, our main contribution involves the proposal of a quality metric, the SUS Consistency Score (CSc), measuring the consistency of a respondent's SUS statements.

We study the effectiveness of the proposed CSc for SUS by conducting a usability study of a recently proposed smartphone security mechanism, PoPLar. Initial findings from a preliminary crowd-sourced usability study of PoPLar had indicated promising results. However, a SUS assessment was not yet performed. The removal of responses based on different quality control thresholds, including CSc, causes significant changes in the obtained SUS score, to the extent of ranking PoPLar differently when compared to a wide range of security proposals for which a SUS score is available. A key implication of this result is that existing SUS scores for all these controls may require revisiting, potentially even revising them upward once the SUS CSc is used as the quality metric to ensure valid responses.

CCS Concepts: • **Security and privacy → Software and application security**; **Usability in security and privacy**.

Additional Key Words and Phrases: Usability, SUS, PoPL, PoPLar, crowd-sourcing

ACM Reference Format:
Yonas Leguesse, Mark Vella, Christian Colombo, and Julio Hernandez-Castro. 2024. A Metric to Assess the Reliability of Crowd-sourced SUS Scores: A Case Study on the PoPLar Authentication Tool. In *The 2024 European Symposium on Usable Security (EuroUSEC 2024), September 30-October 1, 2024, Karlstad, Sweden.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3688459.3688470

1 Introduction

In the evolving landscape of cybersecurity, the emphasis on user experience has become as important as the robustness of security mechanisms themselves. Usability studies are crucial for assessing the overall effectiveness and usability of proposed security mechanisms. However, ensuring the high quality of data collected in these studies is paramount, as low-quality data can lead to misleading conclusions about perceived usability.

Authors' Contact Information: Yonas Leguesse, University of Malta, Msida, Malta, yonas.leguesse.05@um.edu.mt; Mark Vella, University of Malta, Msida, Malta, mark.vella@um.edu.mt; Christian Colombo, University of Malta, Msida, Malta, christian.colombo@um.edu.mt; Julio Hernandez-Castro, Technical University of Madrid, Madrid, Spain, jc.hernandez.castro@upm.es.

The prevalence of bots and low-quality respondents on crowd-sourcing platforms heightens the need for extra attention and robust data validation methods. These issues highlight the complex trade-offs involved in selecting a data collection method and underscore the necessity for researchers to critically assess the potential limitations and benefits of using crowd-sourcing platforms for usability studies. By acknowledging these challenges and employing quality control (QC) strategies [26] to address them, researchers can leverage crowd-sourcing platforms more effectively, potentially enhancing the validity and applicability of usability research in security solutions and beyond.

This work introduces the first self-contained SUS quality metric, the SUS Consistency Score (CSc), which evaluates the reliability of SUS scores gathered through crowd-sourced platforms solely based on SUS responses, without relying on any post-survey quality control. For this exploration, we selected the Proof of Presence and Locality (PoPLar) smartphone authentication tool as a case study. PoPLar is a recently proposed authentication mechanism that uses smartphone sensors and employs a dendrogram puzzle-based approach to enhance usability, allowing for scalable security levels without the need for manual intervention, while offering convenience, privacy, and the avoidance of additional hardware tokens. Initial findings from a preliminary usability assessment indicated promising results regarding its overall usability. However, a SUS assessment had not yet been performed.

During a preliminary test SUS survey on MTurk, we encountered several challenges, such as bot responses and a tendency of participants, either bot or human, to consistently select 'strongly agree' across statements, even when the statements conveyed contradictory sentiments. This indicated bot automation, human carelessness, or a general lack of attention. This observation inspired the concept of consistency scores (CSc) as a response quality metric that quantifies inconsistencies in SUS responses by analysing pairs of sentiment-opposed statements.

Our research involved an extensive analysis of two popular crowd-sourcing platforms, namely Clickworker and Amazon's MTurk, through exploratory surveys. We analysed data obtained from the exploratory surveys using various configurations, such as different platforms and quality controls, to test the hypothesised correlation between CSc and SUS response quality. Measuring the response quality and consistency of the SUS response statements across the exploratory survey configurations was instrumental in validating the CSc quality metric. The results allowed us to assess the quality of the different survey configurations, guiding us to identify a final survey design and configuration that yields high-quality responses.

The final survey's results emphasised the importance of yielding high-quality responses in SUS surveys. We demonstrate that, had we not implemented these quality filters, the reliability of our results would have been compromised by a high number of careless and inconsistent responses. This experience highlighted the challenges of using crowd-sourcing platforms to gather high-quality data, and underscored the necessity of employing robust validation methods to ensure the integrity and reliability of the data collected on crowd-sourcing platforms.

As this paper's key contributions, (i) we critically examine the limitations of crowd-sourcing platforms for usability studies yielding SUS scores, proposing the self-contained SUS Consistency Score (CSc) metric to assess response quality and reliability solely from SUS responses without post-survey quality control, enabling optimized survey design and quality controls, (ii) we perform a usability study of PoPLar, reaffirming its user-centric design, and (iii) we offer a comparative survey of existing usability studies in the security domain that provide SUS scores, showing that the SUS scores at different quality and consistency thresholds rank PoPLar at different positions. We postulate that revisiting all these SUS scores, using CSc as a measure of quality, and applying CSc-based filtering, is necessary in order to evaluate and correct potentially compromised results. Interestingly enough, there is a good chance that these SUS scores may actually increase.

2 Background and Related Work

2.1 System Usability Scale (SUS) and Data Quality Challenges in Crowd-sourcing Platforms

The usability characteristics of any new security solution should be carefully considered. There is often a trade-off between security and usability [15]. In general, usable security does not depend on a single factor but rather on a combination of factors.

The SUS was proposed by John Brooke in 1986 as a means to provide a "quick and dirty" usability scale to evaluate systems [3]. Since its inception, the SUS has been used in numerous usability studies. Ruoti et al. [33] even proposed using the SUS as a standard metric for estimating the relative usability of different authentication schemes. They recommended that new authentication proposals should reach or exceed a baseline SUS score of 68 before receiving serious consideration, a conclusion based on the SUS scores of the seven security tools they investigated.

While not specifically designed for testing usable security, several security tools have employed the SUS in combination with additional questions to further understand specific usability aspects of a security solution (See survey in Section 6.2). A recent review of several publications on the subject of usability in security [18] has identified 14 themes that can significantly impact the usability of security solutions. The themes are intended to provide a comprehensive overview of the various factors at play. While the SUS survey does not cover all 14 themes, it is a good starting point that already addresses several of these aspects.

Brooke's foundational paper [4] discusses the development and rationale behind the SUS, highlighting its intended use and structure. Several studies have reinforced the robustness and reliability of the SUS. Bangor, Kortum, and Miller [1] examined the reliability of the SUS across various studies, demonstrating its reliability and validity in different contexts. Similarly, Sauro et al. [34] investigated the relationship between SUS scores and other usability metrics, providing further evidence for the SUS's robustness. Lewis [19, 20] explored the SUS's application across different user populations and contexts, offering additional insights into its reliability and validity.

Despite the established reliability of the SUS, recent research has highlighted challenges associated with overall data quality from crowd-sourced platforms. Chmielewski et al. [9] noted a decline in data quality since 2018, emphasizing the need for pre-screening and post-survey validations to identify low-quality responses and bots. Similarly, Douglas et al. [10] found that Prolific and CloudResearch generally provided higher quality data compared to MTurk and Qualtrics, underscoring the importance of sophisticated quality control measures. Kennedy et al. [14] identified issues like the use of VPS and fraudulent responses, suggesting stricter attention checks and unique response verification to enhance data quality. However, Zhang et al. [36] found that MTurk, despite being cost-effective, may yield better data quality than commercial panels in terms of completion rates and success in passing manipulation checks. Rouse [31] and Matsuura et al. [25] discussed the necessity of attention and consistency checks and post-survey validation to maintain data integrity in usability studies.

2.2 PoPLar

PoPLar was initially proposed in a recent publication[16] as a solution that makes it possible to securely authenticate financial transactions from a compromised smartphone. The PoPLar implementation, which is sensor-based and centres around a dendrogram design, was highlighted in this study for its effectiveness in countering attacks against real cryptocurrency exchange apps. The PoPLar puzzle solution involves physical interaction with the device, requiring users to tilt their phone to guide a ball through a dendrogram-styled maze to a designated green zone. The maze splits at each intersection, and users can rectify their path by tilting the phone in the opposite direction if they choose incorrectly.

Successfully navigating to and through the green zone completes the current authentication stage, advancing the user towards completing their transaction.

An initial usability assessment yielded encouraging results, indicating that users found it both usable and efficient. However, this initial assessment was somewhat narrow in scope, focusing predominantly on the timing elements and providing only a generic view of the system's usability. The PoPLar puzzle design used in this study includes what is referred to as an intentionality trap. This trap is an additional security feature aimed at mitigating the risk of accidental puzzle solutions, e.g. when the phone is left tilted to one side, allowing for the ball to fall to the far left or far right zones in the puzzle.

3 A Metric for Crowd-sourced SUS Score Reliability

SUS statements are designed in such a way that each statement has an opposite sentiment (positive/negative) when compared to previous or subsequent statements. The SUS score is then calculated by subtracting 1 from the user response for each of the five positive statements (1, 3, 5, 7, 9) and subtracting the user response from 5 for each of the five negative statements (2, 4, 6, 8, 10). The sum of these values, multiplied by 2.5, results in the individual SUS score as a number from 0-100.

With this in mind, an issue identified during initial exploratory surveys involved participants consistently selecting 'strongly agree' to all statements. Because of the way the statements are designed, this clearly indicates inconsistency and/or carelessness. For example, by selecting 'strongly agree' to all statements, a participant is claiming that they strongly agree they would like to use this system frequently (statement 1) while also strongly agreeing that they found the system unnecessarily complex (statement 2). Similarly, they are stating that they strongly agree they found the system easy to use (statement 3) while also strongly agreeing they would need the support of a technical person to use the system (statement 4). These sentiments regarding the tool being assessed are inconsistent, especially considering that the Likert scale allows for a neutral response when the respondent is uncertain of their sentiment.

3.1 Consistency Score (CSc)

Based on this observation, we devised the concept of consistency scores (CSc) to quantify inconsistencies in SUS responses by analysing pairs of sentiment-opposed statements ((1,2), (3,4), ... (9,10)). Consistency is identified when a pair of responses are logically opposing (positive vs. negative), scored at 20% CSc for each pair. Pairs with similar responses to opposed statements are marked as inconsistent, scored at 0% CSc for the pair. Neutrality (a score of 3) in any of the pair's response automatically yields a 20% CSc for that pair, acknowledging neutrality's alignment with both sentiments. The cumulative CSc, ranging from 0% (total inconsistency) to 100% (complete consistency), measures consistency across SUS responses. Below is the mathematical representation for the calculation of the CSc.

Mathematical Representation:

1. **Define Paired Statements:**

Let $P = \{(S_1, S_2), (S_3, S_4), (S_5, S_6), (S_7, S_8), (S_9, S_{10})\}$ be the set of 5 pairs of sentiment-opposed statements derived from the 10 SUS statements, where S_1 is the first SUS statement, S_2 is the second, and so on.

2. **Pair CSc Calculation:** For each pair $(S_i, S_j) \in P$:

$$CSc_{ij} = \begin{cases} 20 & \text{if } S_i \text{ and } S_j \text{ are logically opposing (+ve [4,5] vs -ve [1,2])}, \\ 20 & \text{if } S_i = 3 \text{ or } S_j = 3 \text{ (neutrality)}, \\ 0 & \text{otherwise.} \end{cases}$$

Participant ID	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	SUS	CSc
p1	1	5	1	5	1	5	1	5	1	5	0	100%
p2	5	5	5	5	5	5	5	5	5	5	50	0%
p3	1	1	1	1	1	5	5	5	5	5	55	20%
p4	5	1	4	2	5	1	4	2	3	3	62	100%
p5	5	2	3	2	4	2	4	4	3	3	55	80%
p6	5	1	5	1	5	1	5	1	5	1	100	100%
Average											53.7	66.7%

Table 1. Consistency score examples

3. **Cumulative CSc:** The cumulative CSc is the sum of the scores for each pair:

$$CSc = \sum_{(S_i,S_j)\in P} CSc_{ij}$$

Complete Equation: Combining the steps, the CSc can be expressed as:

$$CSc = \sum_{(S_i,S_j)\in P} \begin{cases} 20 & \text{if } S_i \text{ and } S_j \text{ are logically opposing (+ve vs -ve),} \\ 20 & \text{if } S_i = 3 \text{ or } S_j = 3, \\ 0 & \text{otherwise.} \end{cases}$$

While it is not anticipated that every respondent achieves a perfect 100% CSc, a higher CSc suggests that the respondent devoted time to thoughtfully consider each statement, responding with a level of accuracy and attentiveness indicative of their engagement and careful consideration.

Table 1 shows examples of how the CSc is able to identify responses which appear to be wildly inconsistent. The cells in the table are coloured in green or red depending on whether the pair of statement responses are consistent or not, respectively. In this example, both the second and third participants (*p2* and *p3*) are indicating that they strongly agree with the statement "I felt very confident using the system.", while concurrently stating that they strongly agree with the statement "I needed to learn a lot of things before I could get going with this system.". As a result of these and similar opposing sentiment responses, they end up with a low CSc of 0% and 20%, respectively. On the other hand the first, fourth, and sixth participants (*p1*, *p4*, and *p6*) were consistent in their responses throughout the 10 statements, and all 3 obtained a CSc of 100%. The responses of the fifth participant (*p5*) appear to be consistent for the most part, except for one pair of opposing statements (q7 and q8) to which they responded that they agree with both statements. While this specific pair of responses were inconsistent, it did not significantly impact the overall CSc, with the participant still receiving an overall CSc of 80%. This indicates that for the most part, the responses were consistent.

3.2 Average CSc

Calculating the average CSc allows us to quantify the overall consistency of all of the survey responses in a usability study. The average CSc is computed by calculating the mean of all of the CSc values of the individual participants. Using the values in Table 1 as an example, the average CSc for the group SUS survey responses (*p1* - *p6*) would be:

$$\left(\frac{100 + 0 + 20 + 100 + 80 + 100}{6}\right) = \left(\frac{400}{6}\right) = 66.7\%$$

A higher average CSc indicates that a greater number of respondents provided coherent and reliable answers, thereby enhancing the validity and comparability of the SUS scores obtained from the study. This metric is particularly useful for assessing the quality of responses from crowd-sourcing platforms, where issues such as inattentive or automated

responses can significantly impact data integrity. By calculating the average CSc, researchers can better evaluate and improve the design and implementation of usability surveys, ensuring more accurate and valid results.

3.3 Advantages of CSc

The CSc metric offers significant advantages. Firstly, it does not require any additional quality control questions or statements, as it is derived solely from the SUS statement responses. This means it can be retroactively applied to any previous SUS survey, allowing for the validation of the quality of past survey responses. Additionally, the CSc can be calculated programmatically, making it highly scalable and applicable to larger datasets. Conversely, manual quality control such as the analysis of open-ended text responses, can be a tedious process. Moreover, such manual coding may involving a level of subjectivity, whereas the CSc provides an objective analysis of response quality. These advantages make the CSc a robust and efficient metric for evaluating the quality of SUS responses, providing a mechanism to enhance the reliability and validity of SUS-based usability studies, and their resulting SUS scores.

4 Methodology

To assess the effectiveness of the proposed metric CSc, we conducted a usability study on PoPLar. We analysed the results of exploratory surveys that employed different survey configurations, including the choice of platform and quality controls. This analysis aimed to verify the hypothesised correlation between CSc and SUS response quality. The results were also used to select an optimal survey configuration for the final, larger-scale survey. The results of the final survey were further analysed to highlight the importance of obtaining high-quality responses in SUS surveys, while demonstrating the negative effect low-quality responses can have on the resulting SUS score.

4.1 PoPLar Survey Methodology

Participants were asked to engage with a simulated banking application to execute a sequence of 10 transactions. With each transaction, they encountered a PoPLar challenge, consisting of two levels of dendrogram depth. Data on timing and error rates was collected for subsequent analysis. Upon completion of the transactions, participants were invited to complete a survey assessing their experience with the PoPLar application.

4.1.1 Test Banking App. A simple dummy banking app was developed and published on the Play Store[1]. The idea was to give the participants context of how and when PoPLar challenges are presented through a realistic implementation. The participants were asked to perform 10 transactions each and submit the results to a back-end database hosted on AWS. Through the collected data, we were able to observe timing statistics, error rates, and learnability rates. After the tests, each participant was presented with an unique survey ID to submit together with the survey responses to demonstrate that they completed the tasks successfully. This step acted as an initial quality filter, ensuring that only respondents that actually performed the required tasks were accepted as valid, thus also allowing for the filtering out of bot-responses that randomly populated the survey-ID.

4.1.2 Survey Questions. The survey respondents were asked to respond to the standard SUS statements (See Appendix B), followed by six additional questions that were prepared specifically for PoPLar. Since the SUS statements do not cover all 14 themes of usable security [18], the additional questions were prepared with the objective of validating PoPLar's usability in the context of the 14 usable security themes:

[1]https://play.google.com/store/apps/details?id=com.usability.poplbank

Q1. Was it easy to identify the correct solution for each PoPLar puzzle challenge?
Q2. Do you think that the puzzle design renders the challenge cool and fun?
Q3. Please explain, in your own words, the threat that PoPLar is protecting against. *(Hint: PoPLar protects against a class of attacks known as rDTO (remote device takeover). Through this attack, a cybercriminal can execute tasks on a victim's device without the knowledge of the victim. For example.....)*
Q4. With your understanding of the threat, do you believe that the extra effort required for PoPLar is worth the extra security? If not, explain why.
Q5. Did the presentation of the PoPLar challenge significantly disrupt your primary task (transferring funds)?
Q6. Do you have security or privacy concerns that would stop you from using PoPLar? If so, please explain.

4.1.3 Pre-Screening Quality Controls.

MTurk HIT approval: MTurk provides inbuilt pre-screening quality controls. In this study, we used MTurk's filters to block respondents without a specific number or percentage of successful previously taken tasks, known as Human Intelligence Tasks (HITs). For example, a pre-screening could require that only participants with a HIT acceptance rate (HAR) of 75% are allowed to participate in the survey. This means that if a participant failed more than 25% of previous tasks, they would not be able to participate.

Clickworker Survey ID verification: A survey ID is only presented to the participants after they successfully complete the 10 transactions on the test banking app. Therefore automated bots were not able to complete the required 10 transactions in the test banking application, resulting in their inability to acquire and provide a valid survey ID. Clickworker's inbuilt survey ID verification quality control filter was applied, allowing for the pre-screening of bots by not allowing respondents with invalid survey IDs to submit their responses.

4.1.4 Post-Survey Quality Controls.

Manual Survey ID verification: On MTurk, manual survey ID verification served to filter out bot responses, ensuring that only respondents that actually performed the required tasks were accepted as valid, thus also allowing for the filtering out of bot-responses that randomly populated the survey-ID. Repeated responses were also filtered out through this manual process. Repeated responses refer to responses that were detected to originate from the same respondents through the common unique identifier. In such cases, any of the subsequent submissions from the same respondent were rejected.

Attention Statements: Observations from an initial exploratory survey highlighted the need to be able to identify invalid responses while encouraging valid participation. With this goal in mind, in the subsequent exploratory surveys, we prepared three clear and easy to respond **attention statements (ASt)**, for example:

"This is an attention statement to verify that you are reading the statements. Set this value to 'Strongly agree'. Note, there are more attention statements in this survey."

A failure to correctly answer these simple attention statements casts, in our opinion, serious doubt on the legitimacy and validity of the entirety of the SUS answers provided by the respondent in question. This approach was inspired by Forman et al. [11], who included a single attention statement in the middle of their SUS survey. In the second and third exploratory surveys, we decided to add three attention statements to the survey at the beginning, middle, and end of the SUS statements (see Appendix B). Using three spread out attention statements provides more stringent verification, ensuring attention throughout the survey responses, while also minimising the risk of lucky guesses.

Sensibility of Text Responses: The sensibility of responses was assessed through manual coding, a process involving the qualitative evaluation of open-ended text responses to determine their coherence and relevance. The additional question (Q3) was a suitable candidate to allow for the assessment of response sensibility, since the question was straight forward, did not allow for yes/no responses, and contained a clear hint.

4.2 Validation of CSc and Survey Design

4.2.1 Exploratory Surveys. The exploratory surveys, conducted with the aim of exploring the effectiveness of the proposed quality metric, and evaluating different configurations, were performed on two popular crowd-sourcing platforms: MTurk and Clickworker. Clickworker suited our smartphone-based usability study due to the availability of its app. While other platforms may have also been suitable, Clickworker's features seemed promising for the scope of this study. On MTurk, participants were given $2 (USD) for their participation, while on Clickworker, participants were given 2€ for a 15-minute task. The study was designed to ensure that no sensitive personally identifiable information was collected, and the survey was approved in accordance with our university's research ethics process.

The data from these exploratory surveys, was assessed to:

- validate the effectiveness of the CSc as a quality metric, and
- identify an optimal survey configuration for our final survey which is conducive to:
  - yielding high-quality responses, and
  - minimising the need for manual intervention or assessment.

4.2.2 Validation of CSc. To validate the CSc as a quality metric, we aimed to demonstrate that there is a correlation between CSc scores and response quality. To establish the ground truth in terms of response quality, we first had to differentiate between high and low-quality responses. The following post-survey QC criteria were used to make this distinction:

(1) **Attention Statements (ASt.):** The inability to correctly answer all three attention statements (ASt) indicates a low-quality response.
(2) **Sensibility of Text Responses (sens.):** The provision of a text response that is not considered sensible, determined through manual coding, indicates a low-quality response.

Through manual coding, a response was flagged as being of high-quality if the participant passed all attention statements and provided a sensible text response. These sets of high and low-quality responses were then used as a ground truth to compare and determine the correlation between CSc scores and response quality.

4.2.3 Final Survey Configuration. The final survey configuration was chosen based on the analysis of exploratory surveys, selecting the platform and quality controls that yielded the highest quality responses. This structured approach ensured that both the usability of PoPLar and the effectiveness of the CSc as a quality metric are thoroughly evaluated, providing reliable and valid results for the study.

5 Results

5.1 Exploratory Surveys

Table 2 shows a comparison of the three exploratory surveys (rows ES1, ES2, and ES3), and their respective subsets, comprising a total of 152 responses. Figure 1 shows the relationship between the sets and subsets of the exploratory surveys, including the pre-screening and post-survey quality control filters applied at different stages.

Survey	Platform	Subset	Pre-Screen QC	Post-Survey QC	n	% (/Survey)	SUS (Avg.)	CSc (Avg.)	attn	sens
ES1	MTurk	ES1	N/A	N/A	78	100%	55	42%	N/A	N/A
		$ES1_{rej}$		ID verif. (M)	42	53.8%	52	42%	N/A	N/A
		$ES1_{acc}$			36	46.2%	58	41%	N/A	N/A
ES2	MTurk	ES2	HIT Appr	N/A	36	100%	52	22%	19%	19%
		$ES2_{rej}$		ID verif. (M)	21	58.3%	49	15%	10%	19%
		$ES2_{acc}$			15	41.7%	56	31%	33%	20%
		$ES2_{rej\_inatt}$		- ID verif. (M) - ASt	19	52.8%	49	17%	0%	21%
		$ES2_{rej\_attn}$			2	5.6%	51	0%	100%	0%
		$ES2_{acc\_inatt}$			10	27.8%	55	26%	0%	20%
		$ES2_{acc\_attn}$			5	13.9%	59	40%	100%	20%
ES3	Clickworker	ES3 (acc)	ID verif. (I)	N/A	38	100%	68	69%	61%	63%
		$ES3_{inatt}$		ASt	15	39.5%	60	60%	0%	33%
		$ES3_{attn}$			23	60.5%	73	75%	100%	83%

Table 2. Exploratory Survey Configuration Comparison

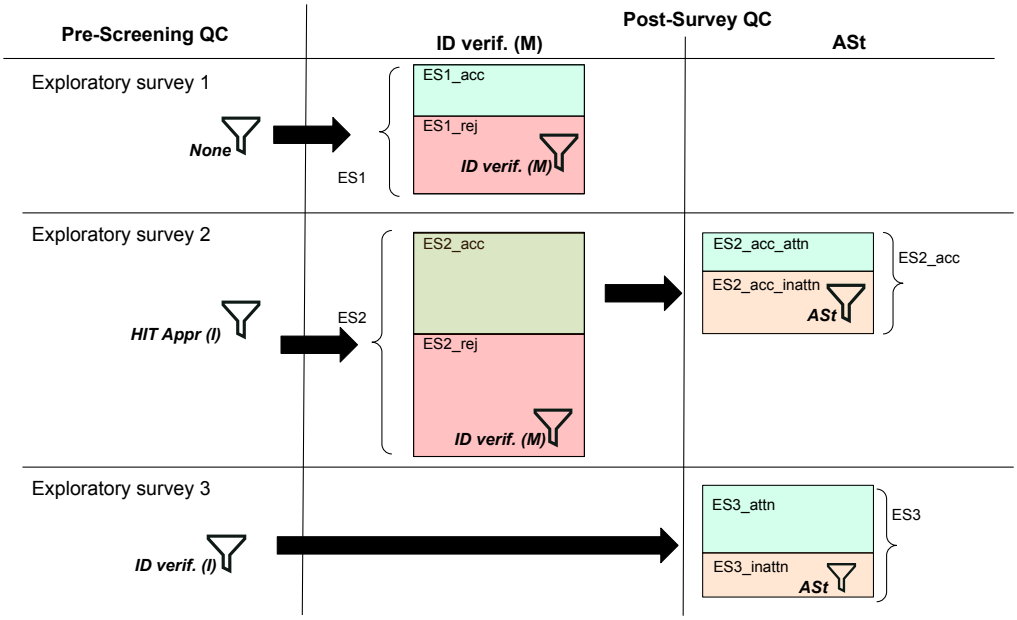

Fig. 1. Exploratory Surveys

5.1.1 Exploratory Survey 1 (ES1). The first exploratory survey was performed on MTurk. This survey did not contain any pre-screening quality controls. Moreover, the additional questions (see Table 7) were not presented as open-ended text, but rather as a Likert scale for each of the 5 questions (Q1-Q2, Q4-Q6) indicating agreement or disagreement with the statements. Q3 was not included in ES1, and was only included in the subsequent exploratory surveys.

- **ES1 (Total):** This set includes all 78 participants from the ES1 survey.
- **$ES1_{rej}$ (Rejected Responses):** These responses were rejected based on a manual survey ID verification (ID verif (M)). This subset represents 53.8% of the total responses, with an average SUS score of 52 and an average CSc of 42%..
- **$ES1_{acc}$ (Accepted Responses):** These responses were accepted based on manual survey ID verification. This subset represents 46.2% of the total responses, with an average SUS score of 58 and an average CSc of 41%.

The prevalence of bot-generated and repeated responses raised significant concerns, highlighting the reality that one cannot simply take responses from crowd-sourcing platforms at face value. This alarming trend underscores the critical need for rigorous validation processes during data collection, as it clearly demonstrates the potential for compromised data integrity in the absence of stringent verification measures.

5.1.2 Exploratory Survey 2 (ES2). The results from ES1 inspired the inclusion of pre-screening QC and the addition of attention statements for post-survey QC (see 4.1.4). Moreover, the additional questions were changed from Likert scale to open-ended text, allowing us to assess the sensibility (sens) of the responses to Q3. ES2 was conducted on MTurk with inbuilt pre-screening QC, that blocked respondents without a specific number or percentage of successful Human Intelligence Tasks (HITs). Initially, we required respondents to have at least 75% of HITs approved. This was subsequently made more stringent by also excluding participants with less than 50 previously approved HITs. Despite these restrictions, post-survey QC still identified several invalid responses through manual survey ID verification, as highlighted below.

- **ES2 (Total):** This set includes all 36 participants from the ES2 survey. The average SUS score is 52 and the average CSc is 22%. Only 19% of respondents provided a sensible text response and 19% passed all 3 attention statements.
- **ES2$_{rej}$ (Rejected Responses):** These responses were rejected based on manual survey ID verification (ID verif (M)). This subset represents 58.3% of the total ES2 responses, with an average SUS score of 49 and an average CSc of 15%. Only 19% of respondents provided a sensible text response and 10% passed all 3 attention statements.
- **ES2$_{acc}$ (Accepted Responses):** These responses were accepted based on a manual survey ID verification. This subset represents 41.7% of the total ES2 responses, with an average SUS score of 56 and an average CSc of 31%. Only 20% of respondents provided a sensible text response and 33% passed all 3 attention statements.
- **ES2$_{rej_inatt}$ (Rejected Inattentive Responses):** These are rejected responses that did not pass all three attention statements (ASt). This subset represents 52.8% of the total ES2 responses, with an average SUS score of 49 and an average CSc of 17%. Only 21% of respondents provided a sensible text response.
- **ES2$_{rej_attn}$ (Rejected Attentive Responses):** These responses passed all three attention statements. This subset represents 5.6% of the total ES2 responses, with an average SUS score of 51 and an average CSc of 0%. None of these respondents provided a sensible text response.
- **ES2$_{acc_inatt}$ (Accepted Inattentive Responses):** These are accepted responses that did not pass all three attention statements (ASt). This subset represents 27.8% of the total ES2 responses and 66.7% of ES2$_{acc}$ responses, with an average SUS score of 55 and an average CSc of 26%. Only 20% of respondents provided a sensible text response.
- **ES2$_{acc_attn}$ (Accepted Attentive Responses):** These responses passed all three attention statements. This subset represents 13.9% of the total ES2 responses and 33.3% of ES2$_{acc}$ responses, with an average SUS score of 59 and an average CSc of 40%. Only 20% of respondents provided a sensible text response.

The pre-screening controls were relatively ineffective, with 58.3% of 36 survey responses failing verification. Only 2 of 21 rejected responses passed attention checks, justifying their rejection. Among 15 accepted responses, 5 passed attention checks, but only 20% of these gave sensible answers, raising concerns about overall response quality.

5.1.3 Exploratory Survey 3 (ES3). The third exploratory survey was conducted on Clickworker. Clickworker's inbuilt survey ID verification pre-screening quality control filter and the attention statements were used as post-survey QC.

- **ES3 (Total):** This set includes all 38 participants from the ES3 survey. The average SUS score is 68, and the average CSc was 69%. Importantly, 63% of the text responses were sensible.

Survey	Platform	Subset	Pre-Screen QC	Post-Survey QC	n	% (/Survey)	SUS (Avg.)	CSc (Avg.)	attn	sens
ESm (ES2 ∪ ES3)	MTurk Clickworker	ESm	ID verif. (I)	- ID verif. (M)	74	100%	60	46%	65%	46%
		ESm_{hq}		- ASt	21	28.4%	76	77%	100%	100%
		ESm_{lq}		- Sens. (M)	53	71.6%	54	34%	17%	24%

Table 3. High vs Low Quality Dataset Comparison

- **$ES3_{inatt}$ (Accepted Inattentive Responses):** These are responses that did not pass all three attention statements. This subset represents 39.5% of the total ES3 responses, with an average SUS score of 60. The average CSc was 60%, and only 33% of responses were sensible.
- **$ES3_{attn}$ (Accepted Attentive Responses):** These responses passed all three attention statements. This subset represents 60.5% of the total ES3 responses, with an average SUS score of 73. The average CSc was 75%, and an impressive 83% of responses were sensible.

Overall, the results of ES3 were more promising than the previous exploratory surveys, showing promise both for the choice of platform and use of attention statements.

5.2 Correlation of CSc and Responses Quality

Table 3 presents data on the subsets of ESm which were used to assess the correlation between CSc and response quality. ESm is a merge (union) of ES2 and ES3 (ES2 ∪ ES3). Responses from the first survey (ES1) were not included in this analysis since this dataset did not contain response sensibility data, which was required in this exercise to establish the ground truth for response quality. The analysis on ESm was further validated by applying the same analysis on the datasets obtained in the final survey (see Table 4), ensuring that the correlation remained consistent even with larger datasets. The data related to the final survey is explained in further detail in section 6.

- **ESm (ES2 ∪ ES3):** This set includes all 74 participants from the ES2 and ES3 surveys, with responses from both MTurk and Clickworker. The average SUS score is 60, and the average CSc is 46%. In this subset, 65% of the respondents managed to pass all 3 attention statements while 46% provided a sensible text response.
- **ESm_{hq} (High-Quality Responses):** This subset includes the 21 responses that passed all three attention statements and provided sensible text responses to Q3. It represents 28.4% of the total ESm responses, with an average SUS score of 76 and an average CSc of 77%.
- **ESm_{lq} (Low-Quality Responses):** This subset includes the 53 responses that either failed an attention statement or did not provide a sensible response. It represents 71.6% of the total ESm responses, with an average SUS score of 54 and an average CSc of 34%. Only 17% of the respondents managed to pass all 3 attention statements and only 24% provided a sensible text response.

The exploratory survey data already indicated a strong correlation between CSc and response quality, with high-quality responses producing an average CSc of 77% and low-quality responses producing an average CSc of 34%. This correlation was also observed in the final survey, where high-quality responses produced an average CSc of 90% and low-quality responses produced an average CSc of 62%. The overall increase in quality from ESm to the final survey was expected, since the final survey used the optimum set of quality controls, as explained in section 5.3.

To test whether high-quality and low-quality responses differ significantly in their CSc scores, we used the Mann-Whitney U test. This test was chosen because it does not assume a normal distribution of CSc values within the groups, with the hypothesis that responses from the high-quality group are more likely to have higher CSc scores.

Comparing ESm_{hq} and ESm_{lq} yielded a p-value < 0.001 and a CLES of 0.8239. Similarly, comparing the final survey results ($Final_{hq}$ and $Final_{lq}$) produced a p-value < 0.001 and a CLES of 0.7685.

The extremely low p-values indicate a significant difference in CSc scores between low-quality and high-quality responses, where high-quality responses are more likely to have higher CSc scores.

Spearman correlation results for ESm, comparing the subsets ESm_{hq} and ESm_{lq}, yielded a p-value < 0.001 and a correlation coefficient of 0.521. The final survey also yielded a p-value < 0.001 and the correlation coefficient was 0.493. These extremely low p-values indicate highly significant correlations between response quality and CSc, with positive correlation coefficients suggesting a meaningful relationship, meaning higher CSc scores are associated with high-quality responses.

The box plot in Figure 2 highlights the difference in distribution of CSc when comparing the low-quality to high-quality responses. The median CSc in ESm_{hq} was 80%, while the median CSc of the high-quality responses in the final survey was 100%. In both the exploratory and final surveys, high-quality responses tended to produce a CSc of 80% or higher. In fact, in the exploratory surveys, 76% of the 21 high-quality responses (ESm_{hq}) produced a CSc of 80% or higher, while only 19% of the 53 low-quality responses (ESm_{lq}) produced a CSc of 80% or higher. This pattern was also observed in the final survey, where an impressive 91% of the 123 high-quality responses produced a CSc of 80% or higher, while only 49% of the 131 low-quality responses produced a CSc of 80% or higher.

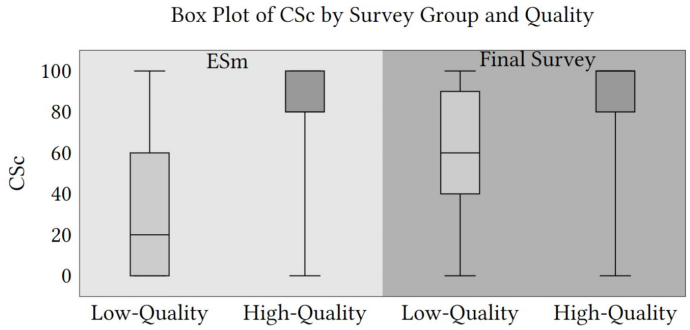

Fig. 2. CSc of Low-Quality and High-Quality Responses in ESm and Final Survey

The strong correlation between CSc and response quality validates CSc as an effective quality metric. The analysis verified that high-quality responses tend to yield high CSc scores, while a low CSc is likely to indicate a low-quality response. Moreover, an average CSc close to, or over, 80% strongly suggests that the survey is made up of high-quality responses.

5.3 Final Survey Configuration

5.3.1 Choice of platform: To ensure a fair platform comparison, we compared ES3 with $ES2_{acc}$ since both subsets consist of responses that passed survey ID verification (see Figure 1). The comparison revealed that Clickworker performed better in terms of participants passing all 3 attention statements and providing sensible text responses. ES3 had 63% sensible text responses and 61% of participants passing all 3 attention statements, while $ES2_{acc}$ had only 20% sensible responses and only 33% of participants passing all 3 attention statements. Additionally, using the CSc, which has now been shown to be a good measure of quality, the average CSc for ES3 was 69%, whereas for $ES2_{acc}$ it was 31%, indicating higher-quality responses in ES3.

To ensure that the CSc scores in the sets were significantly different, a one-sided Mann-Whitney U test was performed. The results showed a p-value < 0.001, confirming the significantly higher CSc, and therefore overall quality, of responses in ES3. Combining these factors led us to choose Clickworker as the platform for the full, larger-scale survey.

5.3.2 Inclusion of attention statements. Comparing $ES3_{attn}$ with $ES3_{inatt}$ validated the inclusion of attention statements as a form of post-survey QC. The high proportion (39.5%) of participants who failed the attention statements and the low amount of sensible text responses (33%) provided by these participants highlighted the prevalence of low-quality responses among participants who failed the attention statements, posing a risk to the resulting SUS score. Comparing the SUS scores of $ES3_{attn}$ (60) with $ES3_{inatt}$ (73) further demonstrates the impact that low-quality responses can have on the resulting SUS score.

5.3.3 Conclusion. Based on the results obtained in the exploratory surveys, for the final survey we decided to:

- Use Clickworker as a crowd-sourcing platform for the remainder of the responses.
- Include attention statements among the SUS statements as a post-survey QC.
- Clearly warn participants that failure to respond correctly to the attention statements will result in the rejection of the survey response.
- Set a response validity threshold and only consider SUS scores as valid for responses with 3/3 correct attention statements.
- Not use the sensibility of the text response as a requirement for validity. This is based on the requirement of finding a configuration conducive to minimising the need for manual intervention and on the high-quality responses obtained from participants that passed all 3 attention statements.
- Compare the resulting average SUS across the final survey responses after applying different post-survey quality control filters to demonstrate the risk that low-quality responses pose to the validity of the resulting SUS score.

6 PoPLar Usability Study Results

6.1 Valid Survey Responses Results

In total, 254 participants responded to the final survey. Out of the 254 responses, 156, or 61.4%, of the respondents passed all three attention statements and were deemed valid. Demographic data for these respondents can be found in Appendix A. The mean SUS score calculated for the 156 valid survey respondents was 75. This SUS score exceeds 68, which Ruoti et al.[33] recommended to consider as a minimum score for new authentication systems before receiving serious consideration. Moreover, according to Bangor et al.'s classification [2], a SUS score of 75 gives PoPLar a 'Good' label rating and places it in the 'Acceptable' range.

Overall, the 156 responses had high consistency scores, with an average CSc of 86%, slightly outperforming the $ES3_{attn}$ exploratory survey which made use of the same quality controls. This slight improvement is possibly due to the inclusion of clear rejection clauses in the final survey. Out of the 156 valid responses, 127 had a CSc that exceeded 80%, and 92 of the 156 responses achieved 100% CSc, underscoring the response reliability and validating the use of attention statements to eliminate low-quality responses. While we did not set the sensibility of text responses as a criteria for validity, manual coding of the sensible text responses was already performed to correlate the CSc with high-quality responses, as described in section 5.2. Using this data, we could see that 79% of the 156 valid responses provided sensible text responses, indicating that the selected survey configuration was successful in yielding high-quality responses.

In addition to the SUS statements, respondents were asked questions related to the PoPLar tool (see Section 4.1.2). Participants generally found it easier to solve PoPLar puzzles on a second attempt, indicating a positive learning

Survey	Platform	Subset	Pre-Screen QC	Post-Surv QC	n	% (/Survey)	SUS (Avg.)	CSc (Avg.)	attn	sens	hq
Final	Clickworker	Final	ID verif. (I)	N/A	254	100%	68	76%	61%	60%	48%
		Final_{attn}		ASt	156	61.4%	75	86%	100%	79%	79%
		Final_{inatt}			98	38.6%	56	60%	0%	30%	0%
		Final_{hq}		- ASt - Sens. (M)	123	48.4%	78	90%	100%	100%	100%
		Final_{lq}			131	51.6%	58	62%	25%	22%	0%
		Final_{80p}		CSc	175	68.9%	73	93%	83%	79%	64%
		Final_{u80}			79	31.1%	54	37%	57%	30%	14%
		Final_{attn_80p}		- CSc - ASt	127	50%	78	94%	100%	88%	88%

Table 4. Final Survey Subset Comparison

curve. Most received the puzzle design positively, though some suggested varied difficulty levels to maintain interest. Respondents considered the added security worth the extra effort, with suggestions to balance challenge frequency with user convenience. Most did not find the PoPLar challenge disruptive to their tasks and expressed trust in PoPLar's ability to enhance transaction security without compromising privacy.

The mean time-to-solve (TTS) for a challenge was 23 seconds, with a median of 19 seconds and a standard deviation of 16 seconds. For a single level, the mean TTS was 12 seconds, the median was 9 seconds, and the standard deviation was 10 seconds. The failure rate was low at 2.64%, and the data indicated high learnability, with the mean TTS decreasing from 23 seconds on the first attempt to 12.6 seconds on the second, and stabilizing to just under 10 seconds for subsequent attempts.

6.2 Impact of Quality Controls on SUS score

To demonstrate the impact that the quality of SUS responses can have on the final SUS score, we analysed the results of different subsets from the final PoPLar survey, grouped by different types of post-survey QC (see Table 4). We then compared the SUS scores of these subsets to SUS scores of 51 security tools obtained from 18 different usability studies.

6.2.1 Final Survey Subsets.

- **Final (Total):** This set includes all 254 valid and invalid participants from the final survey. The resulting average SUS score is 68, with an average CSc of 76%. In this subset, 60% of the text responses were considered sensible.
- **$\text{Final}_{\text{attn}}$ (Valid Attentive Responses):** These responses passed all three attention statements and were therefore considered valid. This subset represents 61.4% of the total responses, with an average SUS score of 75. The valid responses provided an impressive average CSc of 86%, and 79% provided sensible text responses.
- **$\text{Final}_{\text{inatt}}$ (Rejected Inattentive Responses):** These are responses that did not pass all three attention statements. This subset represents 38.6% of the total responses, with an average SUS score of 56, and an average CSc of 60%. Only 30% of these responses provided a sensible text response indicating a large amount of low-quality responses.
- **Final_{hq} (High-Quality Responses):** These are responses that passed all three attention statements and also provided sensible text responses. This subset represents 48.4% of the total responses, and 78.8% of the accepted responses. These responses had an average SUS score of 78 with an average CSc of 90%.
- **Final_{lq} (Low-Quality Responses):** These are responses that either did not pass all three attention statements or did not provide sensible text responses. This subset represents 51.6% of the total responses, and obtained an average SUS score of 58. The average CSc was 62%, and only 22% provided sensible text responses.
- **$\text{Final}_{\text{80p}}$ (80%-Plus CSc):** These are responses from the final survey, that obtained a CSc of 80% or over, meaning the CSc was the only post-survey QC filter applied. This subset represents 68.9% of the total responses, and obtained

an average SUS score of 73. An impressive 83% of respondents passed all 3 attention statements and 79% provided sensible text responses. In total 64% provided high-quality responses.

- **$\text{Final}_{\text{u80}}$ (CSc Under 80%):** These are responses from the final survey, that obtained a CSc lower than 80%. This subset represents 31.1% of the total responses, and obtained an average SUS score of 54. In this subset, 57% of respondents passed all 3 attention statements and only 30% provided sensible text responses. In total only 14% provided high-quality responses.
- **$\text{Final}_{\text{attn\_80p}}$ (Valid Attentive Responses with 80%-Plus CSc):** These represent the subset of $\text{Final}_{\text{attn}}$ that obtained a CSc of 80% or over, meaning that both ASt and the CSc were used as post-survey QC. This subset represents 50% of the total responses, and obtained an average SUS score of 78, matching the SUS score obtained in the Final_{hq}. In this subset, 88% of these respondents provided sensible text responses, rendering them high-quality responses by definition.

6.2.2 SUS scores from existing usability studies. We reviewed 18 different usability studies that yielded SUS scores for a combination of 51 security tools. Table 5 lists the 51 tools and their respective SUS scores, as well as the various SUS scores obtained for PoPLar in each of the survey subsets listed in Table 4. The table includes information on the survey setting (Survey), the resulting SUS score, the number of survey participants (n), the quality control measures mentioned in the paper (QC), and the average CSc (ACSc). Since the CSc was introduced in this research, only the PoPLar results currently include this metric. However, given that the CSc can be applied retroactively to existing SUS surveys, provided the raw survey data, one could also compute the average CSc scores for previous studies.

Our review indicates that quality control practices, particularly with regard to the evaluation of SUS survey responses, are either absent or not explicitly mentioned in the majority of the listed usability studies. Only one study [32] dedicated a section to quality controls, however, this section focused more on user issues when handling the solution (SH) rather than the quality of SUS survey responses. Other papers only mentioned survey quality controls aspects in passing. Interestingly, a survey [29] conducted on a crowd-sourcing platform mentioned the use of MTurk qualifications and geolocation checks to filter for bots (B), but our analysis in section 5.1 questions the effectiveness of using only MTurk's in-built qualification filters against bot responses. Another crowd-sourced survey [11] implemented attention statements (ASt) to filter out inattentive participants. However, the very low failure rate among participants suggests limited effectiveness. A significant portion of these studies were conducted in university (Uni) settings, which, while convenient, present a number of limitations, including: a relatively homogeneous population, the often limited sample size (n), the time required, and the associated costs. Similar limitations apply for lab (Lab), recruitment firm (Rec), and organisational employee (Empl) settings. On the other hand, crowd-sourced (CS) or social media (SM) surveys offer a quicker, typically less expensive alternative, providing easy access to a larger and more heterogeneous population [22].

6.2.3 SUS score comparison. Table 5 also shows how the SUS scores of the different subsets from the final survey rank in comparison to the other security proposals (Rank). The rank ranges from 1 to 59, where 8 of these ranks are assigned to each of the subsets highlighted in Table 4. By comparing these subsets, we can observe the variation in SUS scores and position rankings among the 59 SUS scores, based on different quality controls.

For instance, the three subsets Final_{lq}, Final_{inatt}, and Final_{u80} obtained a SUS score of 58, 56, and 54 respectively, all falling below Ruoti et al.'s proposed baseline SUS score of 68 [33], ranking PoPLar at 52nd, 54th, and 55th place respectively out of the 59 SUS scores. Commonalities among these three subsets include a low average CSc (well below 80%), and poor results in terms of attention statements and/or text response sensibility, all being indicators of low-quality responses.

Paper	Security Solution	Rank	SUS	n	Survey	QC	ACSc
P13[35]	Face+Voice	59	46	30	Empl	—	—
P15[17]	BoD Shapes 1	58	47.70	12	Uni	—	—
P13[35]	Gesture+voice	57	50	30	Empl	—	—
P1[32]	SSo: Hatchet	56	53.50	18	Uni	SH	—
P19	PoPLar: (Final$_{u80}$)	55	54	79	CS	Table 4	37%
P19	PoPLar: (Final$_{inatt}$)	54	56	98	CS	Table 4	60%
P1[32]	QR: WebTicket	53	57.90	25	Uni	SH	—
P19	PoPLar: (Final$_{lq}$)	52	58	131	CS	Table 4	62%
P11[13]	PATTERN	51	60	30	Uni	—	—
P1[32]	SSo: SAW	50	61	48	Uni	SH	—
P3[28]	WebAuthn	49	63.50	10	Rec	—	—
P7[12]	GR-password	48	64	102	Misc.	—	—
P13[35]	Voice	47	66	30	Emp.	—	—
P4[29]	Neo	16	66.60	44	CS	B	—
[33]	Baseline for consideration	N/A	68	N/A	—	—	—
P19	PoPLar: (Final)	45	68	254	CS	Table 4	76%
P5[21]	SmartThings	44	68.08	30	Uni	—	—
P12[7]	Hold & Sign	43	68.33	30	Uni	—	—
P17[5]	UsPi	42	68.75	16	Uni	—	—
P11[13]	PIN	41	69.50	30	Uni	—	—
P1[32]	SSo: FB	40	71.40	24	Uni	SH	—
P7[12]	Fingerprint	39	71.50	100	Misc.	—	—
P8[24]	text-passwords	38	71.77	48	SM/Uni	—	—
P1[32]	SSo: Mozilla	37	71.80	24	Uni	SH	—
P1[32]	SSo: Google	36	72	54	Uni	SH	—
P5[21]	P2Auth: knob	35	72.92	30	Uni	—	—
P19	PoPLar: (Final$_{80p}$)	34	73	175	CS	Table 4	93%
P2[30]	U2F key	33	73.10	72	Uni	—	—
P5[21]	P2Auth: swiping	32	73.21	30	Uni	—	—
P6[11]	DPatt	30	73.21	634	CS	ASt	—
P10[8]	DialerAuth	30	73.29	97	CS	—	—
P15[17]	BoD Shapes 2	29	73.50	12	Uni	—	—
P5[21]	P2Auth: button	28	74.16	30	Uni	—	—
P2[30]	SS	27	75	72	Uni	—	—
P13[35]	Face	26	75	30	Empl	—	—
P19	PoPLar (Final$_{attn}$)	25	75	156	CS	Table 4	86%
P9[6]	AnswerAuth	24	75.11	85	CS	—	—
P1[32]	QR: Snap2Pass	23	75.70	55	Uni	SH	—
P13[35]	Gesture	22	77	30	Empl	—	—
P7[12]	Pattern-lock	21	78	101	Misc.	—	—
P13[35]	Password	20	78	30	Empl	—	—
P14[37]	PIN	19	78	20	Lab	—	—
P19	PoPLar: (Final$_{attn_80p}$)	18	78	127	CS	Table 4	94%
P19	PoPLar: (Final$_{hq}$)	17	78	123	CS	Table 4	90%
P2[30]	Pre gen-codes	16	80.20	72	Uni	—	—
P15[17]	BoD Taps 1	15	80.40	12	Uni	—	—
P16[23]	fingerprint	14	80.60	35	Uni	—	—
P2[30]	Push	13	81	72	Uni	—	—
P8[24]	Yubico Key	12	81.74	46	SM/Uni	—	—
P7[12]	4-digit PIN	11	82	100	Misc.	—	—
P15[17]	BoD Taps 2	10	82.10	12	Uni	—	—
P18[27]	TapMeIn	9	83	41	SM	—	—
P2[30]	TOTP	8	83.10	72	Uni	—	—
P17[5]	UsPa	7	84.10	16	Uni	—	—
P14[37]	RiskCog	6	84.50	20	Lab	—	—
P4[29]	Password	5	88.60	66	CS	B	—
P17[5]	Fiduciary tag	4	90.20	16	Uni	—	—
P2[30]	Password	3	92.50	72	Uni	—	—
P17[5]	TaPi	2	92.50	16	Uni	—	—
P11[13]	GRA-PIN	1	94	30	Uni	—	—

Table 5. SUS score surveys for various security proposals (Sorted by SUS score. Red=This paper. Green = Baseline for consideration)

The Final set illustrates the results had we not performed any post-survey QC in our survey. The resulting SUS score of 68 places PoPLar dangerously close to the proposed baseline score of 68, also ranking PoPLar at the 43rd position. However, as soon as we start applying post-survey QC, we can see an improvement in the SUS score and overall ranking. Interestingly, using just the CSc as post-survey QC filter (Final_{80p}), we see that the SUS score increases to 73, placing it above the baseline of 68 with a 34th position ranking. Moreover, this filtering resulted in an increase in response quality with 83% of respondents passing all 3 attention statements and 79% providing sensible text responses. In total 64% of Final_{80p} provided high-quality responses. These results show that the CSc-based filtering could be used as a post-survey QC filter to increase the quality and reliability of responses and the resulting SUS score.

Similarly, using the attention statements as quality controls, i.e. removing the Final_{inatt}, increased the SUS score, providing a score of 75 with a position ranking of 25th place. Interestingly, we observe a further increase in SUS if we apply even more stringent quality controls. By adding the CSc as post-survey QC filter to Final_{attn}, the resulting SUS score for Final_{attn_80p} is 78, ranking it at 17th/18th place. Similarly, by only choosing the high-quality responses from the entire Final set of 254 responses, the resulting SUS score is also 78, ranking it at an equal 17th/18th place.

7 Conclusion and Future Work

This work highlighted the importance of implementing robust quality controls in usability studies conducted on crowd-sourcing platforms to obtain valid and reliable System Usability Scale (SUS) scores. A series of exploratory surveys performed using two leading crowd-sourcing survey platforms and a recently proposed smartphone authentication mechanism revealed substantial variations in SUS scores with respect to varying degrees of quality controls. This led to the formulation of a consistency score (CSc) for SUS studies. Using quality controls, including the CSc as a quality metric, we were able to compare the exploratory surveys to identify a configuration that retained the largest percentage of high-quality responses and achieved a high average CSc. The results demonstrate the effectiveness of the quality controls in enhancing data quality while also validating the use of CSc as a quality metric.

These insights highlight the need for careful thought in survey design and configuration to ensure trustworthy SUS scores. By carefully selecting quality controls, researchers can enhance the overall quality of usability studies and achieve more reliable and comparable SUS scores. The final survey configuration, i.e. using Clickworker and attention statements, validated this approach with an impressive average CSc of 86%, further reinforcing the value of these methodological improvements.

The PoPLar case study provided insightful findings, showing that depending on the subset of survey responses, the same tool received a SUS score ranging from 56 to 78. This underscores the impact that response quality has on SUS surveys, where unreliable responses can result in inflated or deflated SUS scores, potentially presenting an incorrect view of a tool's usability. Unlike this study, past SUS-based usability studies identified in this work overlooked such consistency and quality metrics to verify the validity of responses. Future work should look into revisiting prior SUS-survey data to retroactively calculate the consistency scores and possibly apply CSc-based filtering, which could greatly benefit the research community by ensuring that SUS scores are more comparable and reliable. This could potentially lead to the reconsideration of systems that may have been previously overlooked due to underreported SUS scores

Appendices

A Final Survey Demographics

		#	%
Total		156	100
Gender	**Female**	51	33
	Male	105	67
Generation	**Gen-Z**	28	18
	Millennial	100	64
	Gen-X	22	14
	Boomer	4	3
	Not Provided	2	1
Computer Skill	**Expert**	30	19
	Advanced	75	48
	Intermediate	38	24
	Basic	13	8

Table 6. Survey demographics

B SUS Statements with Attention Statements

SUS Statement	**Strongly Disagree**				**Strongly Agree**
	1	**2**	**3**	**4**	**5**
ASt1: ..Set this value to 'Strongly agree'..					
I think that I would like to use this system frequently.					
I found the system unnecessarily complex.					
I thought the system was easy to use.					
I think that I would need the support of a technical person to be able to use this system.					
I found the various functions in this system were well integrated.					
ASt2: ..Set this value to 2..					
I thought there was too much inconsistency in this system.					
I would imagine that most people would learn to use this system very quickly.					
I found the system very cumbersome to use.					
I felt very confident using the system.					
I needed to learn a lot of things before I could get going with this system.					
ASt3: ..Set this value to 'Strongly disagree'..					

Table 7. System Usability Scale w/ ASt

References

[1] Aaron Bangor, Philip Kortum, and James Miller. 2008. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction* 24, 6 (2008), 574–594.

[2] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[3] John Brooke. 1986. System usability scale (SUS): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd* 43 (1986), 1–7.

[4] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[5] Anders Bruun, Kenneth Jensen, and Dianna Kristensen. 2014. Usability of single-and multi-factor authentication methods on tabletops: a comparative study. In *Human-Centered Software Engineering: 5th IFIP WG 13.2 International Conference, HCSE 2014, Paderborn, Germany, September 16-18, 2014. Proceedings 5*. Springer, 299–306.

[6] Attaullah Buriro, Bruno Crispo, and Mauro Conti. 2019. AnswerAuth: A bimodal behavioral biometric-based user authentication scheme for smartphones. *Journal of information security and applications* 44 (2019), 89–103.

[7] Attaullah Buriro, Bruno Crispo, Filippo Delfrari, and Konrad Wrona. 2016. Hold and sign: A novel behavioral biometrics for smartphone user authentication. In *2016 IEEE security and privacy workshops (SPW)*. IEEE, 276–285.

[8] Attaullah Buriro, Bruno Crispo, Sandeep Gupta, and Filippo Del Frari. 2018. Dialerauth: A motion-assisted touch-based smartphone user authentication scheme. In *Proceedings of the eighth ACM conference on data and application security and privacy*. 267–276.

[9] Michael Chmielewski and Sarah C Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473.

[10] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one* 18, 3 (2023), e0279720.

[11] Tim Forman and Adam Aviv. 2020. Double patterns: a usable solution to increase the security of android unlock patterns. In *Annual Computer Security Applications Conference*. 219–233.

[12] Kelly Grindrod, Hassan Khan, Urs Hengartner, Stephanie Ong, Alexander G Logan, Daniel Vogel, Robert Gebotys, and Jilan Yang. 2018. Evaluating authentication options for mobile health applications in younger and older adults. *PloS one* 13, 1 (2018), e0189048.

[13] Nabeela Kausar, Ikram Ud Din, Mudassar Ali Khan, Ahmad Almogren, and Byung-Seo Kim. 2022. GRA-PIN: A Graphical and PIN-Based Hybrid Authentication Approach for Smart Devices. *Sensors* 22, 4 (2022), 1349.

[14] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629.

[15] Oksana Kulyk, Stephan Neumann, Jurlind Budurushi, and Melanie Volkamer. 2017. Nothing comes for free: How much usability can you sacrifice for security? *IEEE Security & Privacy* 15, 3 (2017), 24–29.

[16] Yonas Leguesse, Christian Colombo, Mark Vella, and Julio Hernandez-Castro. 2021. PoPL: Proof-of-Presence and Locality, or How to Secure Financial Transactions on Your Smartphone. *IEEE Access* 9 (2021), 168600–168612.

[17] Luis A Leiva and Alejandro Català. 2014. BoD taps: an improved back-of-device authentication technique on smartphones. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. 63–66.

[18] Markus Lennartsson, Joakim Kävrestad, and Marcus Nohlberg. 2021. Exploring the meaning of usable security–a literature review. *Information & Computer Security* 29, 4 (2021), 647–663.

[19] James R Lewis and Jeff Sauro. 2018. Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability metric. *International Journal of Human-Computer Interaction* 34, 6 (2018), 561–567.

[20] James R Lewis and Jeff Sauro. 2019. Item Benchmarks for the System Usability Scale. *Journal of Usability Studies* 14, 3 (2019), 149–173.

[21] Xiaopeng Li, Fengyao Yan, Fei Zuo, Qiang Zeng, and Lannan Luo. 2019. Touch well before use: Intuitive and secure authentication for iot devices. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[22] Di Liu, Randolph G Bias, Matthew Lease, and Rebecca Kuipers. 2012. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology* 49, 1 (2012), 1–10.

[23] Sajaad Ahmed Lone and AH Mir. 2022. Smartphone-based Biometric Authentication Scheme for Access Control Management in Client-server Environment. (2022).

[24] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. 2020. Is FIDO2 the kingslayer of user authentication? A comparative usability study of FIDO2 passwordless authentication. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 268–285.

[25] Tenga Matsuura, Ayako A Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. 2021. Careless participants are essential for our phishing study: Understanding the impact of screening methods. In *Proceedings of the 2021 European Symposium on Usable Security*. 36–47.

[26] Christopher D Mellinger and Thomas A Hanson. 2020. Methodological considerations for survey research: Validity, reliability, and quantitative analysis. *Linguistica Antverpiensia, New Series–Themes in Translation Studies* 19 (2020).

[27] Toan Nguyen and Nasir Memon. 2018. Tap-based user authentication for smartwatches. *Computers & Security* 78 (2018), 174–186.

[28] Wataru Oogami, Hidehito Gomi, Shuji Yamaguchi, Shota Yamanaka, and Tatsuru Higurashi. 2020. Observation study on usability challenges for fingerprint authentication using WebAuthn-enabled android smartphones. *Age* 20 (2020), 29.

[29] Kentrell Owens, Olabode Anise, Amanda Krauss, and Blase Ur. 2021. User perceptions of the usability and security of smartphones as {FIDO2} roaming authenticators. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 57–76.

[30] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. 2019. A usability study of five {two-factor} authentication methods. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. 357–370.

[31] Steven V Rouse. 2019. Reliability of MTurk data from masters and workers. *Journal of Individual Differences* (2019).

[32] Scott Ruoti, Brent Roberts, and Kent Seamons. 2015. Authentication melee: A usability analysis of seven web authentication systems. In *Proceedings of the 24th international conference on world wide web*. 916–926.

[33] Scott Ruoti and Kent Seamons. 2016. Standard metrics and scenarios for usable authentication. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*.

[34] Jeff Sauro and James R Lewis. 2011. Measuring usability with the System Usability Scale (SUS). *J Usability Studies* 4, 3 (2011), 194–198.

[35] Shari Trewin, Cal Swart, Larry Koved, Jacquelyn Martino, Kapil Singh, and Shay Ben-David. 2012. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*. 159–168.

[36] Bingbing Zhang and Sherice Gearhart. 2020. Collecting online survey data: A comparison of data quality among a commercial panel & MTurk. *Survey Practice* 13, 1 (2020), 1–10.

[37] Tiantian Zhu, Zhengyang Qu, Haitao Xu, Jingsi Zhang, Zhengyue Shao, Yan Chen, Sandeep Prabhakar, and Jianfeng Yang. 2019. RiskCog: Unobtrusive real-time user authentication on mobile devices in the wild. *IEEE Transactions on Mobile Computing* 19, 2 (2019), 466–483.