Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Digital Data Collection among Low ICT-Literate Rural Communities: A Case Study using Google Forms via Smartphones

Wan Hussain Wan Ishak[1], Fadhilah Mat Yamin[2], Risyawati Mohamed Ismail[3], Mastora Mustafar[3], Siti Zakiah Abu Bakar[3]

[1]*School of Computing, Universiti Utara Malaysia, Malaysia*
[2]*Knowledge Science Research Lab, School of Technology Management and Logistics, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*
[3]*School of Technology Management and Logistics, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*

ARTICLE INFO

Article history:
Received 30 June 2025
Revised 05 July 2025
Accepted 07 July 2025
Available online 07 August 2025

Keywords:
Google Forms
rural communities
low ICT literacy
smartphone surveys
data quality
digital data collection
rural ICT access

Please cite this article in IEEE style as:
W. H. Wan Ishak, F. Yamin, R. M. Ismail, M. Mustafar, and S. Z. Abu Bakar, "Digital Data Collection among Low ICT-Literate Rural Communities: A Case Study using Google Forms via Smartphones", Data Science Insights, pp. 1–5, Aug. 2025.

ABSTRACT

This study investigates the use of Google Forms as a digital tool for daily livestock monitoring among rural, low ICT-literate chicken farmers in Malaysia. A total of 198 responses were collected via smartphones through WhatsApp-distributed forms, allowing participants to self-report poultry conditions while reducing the need for frequent site visits. While the approach proved accessible and cost-effective, analysis revealed significant data quality issues, including inconsistent data entry (e.g., mixed numeric and textual values), unstructured categorical responses, duplicate submissions, ambiguous placeholder values, and the absence of unique identifiers. These challenges limited the reliability and usability of the dataset for meaningful analysis. To address these issues, the study recommends implementing structured input fields, validation rules, unique respondent IDs, and user training materials tailored to low digital literacy. This paper highlights both the potential and pitfalls of digital self-reporting tools in underserved rural contexts and provides practical recommendations for improving data quality in similar monitoring efforts. The findings offer valuable guidance for researchers and practitioners designing data collection systems in constrained environments.

Correspondence:
Wan Hussain Wan Ishak
School of Computing, Universiti Utara Malaysia, Malaysia
hussain@uum.edu.my

1. Introduction

The rise of digital technology has transformed the landscape of data collection, offering researchers a range of tools that are faster, more scalable, and less resource-intensive compared to traditional methods. Platforms such as Google Forms (https://docs.google.com/forms/), JotForm (https://www.jotform.com/), and ODK (https://getodk.org/) have become increasingly popular, particularly for research in remote or resource-constrained settings. These tools reduce the need for manual data entry and enable researchers to receive submissions in real-time, enhancing the speed of analysis and decision-making [1]. However, the adoption of such tools must be carefully aligned with the context in which data is being collected, especially in environments characterized by low digital literacy and limited internet connectivity.

Selecting the appropriate data collection method is not a purely technical decision but a strategic one. Taherdoost [2] emphasizes that the success of a research project depends significantly on the appropriateness of its data collection technique, which should be guided by the nature of the research question, the profile of the respondents, and the physical and technological environment. In rural or underdeveloped areas, where respondents may lack exposure to formal ICT tools, digital surveys must be designed with a strong emphasis on usability, clarity, and simplicity. This requirement becomes even more critical when researchers have minimal physical access to the participants and rely entirely on remote digital tools.

Despite the benefits of digital surveys, they present unique challenges in low-literacy and low-connectivity settings. As Fitzgerald and FitzGibbon [3] argue, while digital data collection can empower rural communities by offering more flexible participation, it also risks misunderstanding or disengagement when tools are not adapted to local capacities. In many cases, respondents unfamiliar with digital forms may struggle with input fields, submit incomplete or inconsistent responses, or misinterpret the survey's intent. These challenges are magnified when data collection is conducted via smartphones—a device often limited by screen size, typing difficulty, or restricted data plans. This is similarly observed in the context of customer satisfaction towards self-service technologies in Malaysia, where users' interaction with digital interfaces is crucial for the system's success [4].

Data quality is another critical consideration in such contexts. According to Hassenstein and Vanella [5], high-quality data must meet criteria such as accuracy, completeness, consistency, and interpretability. These dimensions are frequently compromised in digital surveys involving low-literate users. For example, free-text fields without input validation often receive vague, irrelevant, or inconsistent responses. Similarly, numeric fields may contain placeholder text, incorrect formats, or mixed units. Without appropriate design strategies and quality control mechanisms, such issues can undermine the reliability of the collected dataset and affect the validity of any subsequent analysis. The importance of ensuring data quality is underscored by research that has examined large datasets, where quality issues like redundancy and inconsistencies were present [6].

The complexity of managing and assessing data quality increases when responses are submitted repeatedly without unique identifiers or when duplicate records appear due to confusion or submission errors. Shanmugam et al. [7] highlight the particular challenges of handling such inconsistencies in data-intensive environments, noting that even small-scale projects must anticipate issues such as data redundancy, missing values, and irregular time-stamping. These are not merely technical problems but symptoms of a larger gap between technological solutions and their practical use among marginalized populations. For researchers working in the field, especially in low-income rural communities, these concerns must be addressed from the outset to ensure the data collected is both meaningful and actionable.

This article presents a case study of digital data collection conducted among rural poultry farmers in Malaysia. The study involved administering a survey via Google Forms to respondents with low ICT literacy, all of whom completed the form using smartphones under conditions of limited internet access. The paper identifies key data quality issues arising from this context, including inconsistent entries, lack of standardization, absence of unique identifiers, and potential respondent misunderstanding. Drawing from the findings, it offers concrete recommendations for researchers aiming to implement effective and context-sensitive digital data collection strategies in similar environments.

2. Literature Review

Data quality has become an essential focus in both academic and practical domains due to its direct impact on the validity and reliability of decision-making processes. As organizations and researchers increasingly rely on data to inform strategies and policies, ensuring that data are accurate, complete, and timely is more important than ever. According to Hassenstein and Vanella [5], data quality comprises several dimensions—such as accuracy, completeness, consistency, validity, timeliness, and uniqueness—which together determine whether data are fit for their intended use. However, achieving and maintaining these qualities remains a persistent challenge, particularly in environments where data are collected across diverse formats and under varying levels of technological access.

The method of data collection significantly affects the quality of data gathered. Taherdoost [2] outlines a range of both traditional and digital data collection techniques, each with distinct advantages and limitations. Traditional methods, such as face-to-face interviews and paper-based surveys, often offer richer contextual insights but are time-consuming and prone to manual errors. On the other hand, digital tools—ranging from online forms to mobile-based data collection apps—offer speed, scalability, and cost-efficiency. Singh and Burgess [1] highlight that electronic data collection, while efficient, may introduce new forms of error, such as incomplete submissions, duplicate entries, or misinterpretation of questions due to interface limitations. Fitzgerald and FitzGibbon [3] further caution that in low-developed countries (LDCs), digital data collection may unintentionally produce biased or misleading data, especially where respondents possess limited technological literacy or access to reliable internet. These findings underscore the importance of aligning data collection methods with the socio-technical context of the target population.

Recent research trends show a growing interest in the challenges of data quality within the context of big data. Cai and Zhu [8] note that the sheer volume, velocity, and variety of data in the big data era exacerbate existing data quality problems. Traditional data validation methods are often inadequate for dealing with unstructured data or real-time data streams. Shanmugam et al. [7] similarly argue that managing data quality at scale requires new paradigms, as current assessment frameworks lack the capacity to accommodate big data's complexity and speed. The limitations of existing approaches are also highlighted by Jaya et al [9], whose systematic review reveals that much of the research on data quality remains fragmented and technically oriented, with insufficient attention given to contextual and organizational factors that may influence data quality.

The practical application of data quality assessment spans multiple domains. Dravis [10] provides a pragmatic framework for developing a data quality strategy, emphasizing the need for clearly defined standards, designated

data stewards, and continuous monitoring and auditing. In a similar vein, Li et al [11] propose a systematic model for addressing data quality issues through a combination of error prevention, detection, and correction techniques. These frameworks are useful for organizations aiming to institutionalize data quality assurance practices, yet their implementation can be hampered by technical constraints and a lack of qualified personnel, particularly in resource-limited settings.

Several limitations continue to hinder advancements in data quality management. One significant challenge is scalability. Methods that are effective at small or moderate data volumes may falter when applied to high-frequency data streams or massive datasets. Another issue lies in automation. Many quality checks are still performed manually or semi-manually, delaying timely interventions. Additionally, existing models often assume a high degree of user digital literacy, which may not hold true in various practical scenarios, especially when collecting data from marginalized or digitally underserved communities [3].

Looking ahead, future research in data quality should focus on bridging the gap between technical sophistication and user accessibility. There is a need for intelligent data quality tools that leverage machine learning and artificial intelligence to detect anomalies, assess context, and adaptively validate data inputs in real time. Developing domain-specific quality standards and interoperability protocols can also promote consistency across datasets and systems. Moreover, more studies are needed to explore how data collection tools can be designed to accommodate varying levels of user capability and infrastructural limitations, as emphasized by [2]. Integrating human-centered design principles and participatory approaches into the development of data systems could lead to more inclusive, accurate, and contextually appropriate data collection and quality assurance methods.

3. Research Methods

This study employed a structured digital questionnaire as part of a broader effort to monitor small-scale poultry farming activities in a rural Malaysian community. The questionnaire was developed using Google Forms and disseminated via WhatsApp, a platform familiar and accessible to the target population. The primary purpose of the form was to enable real-time, remote monitoring of chicken rearing conditions without requiring frequent physical visits to the field. By allowing participants to submit daily updates on their chickens, this approach facilitated ongoing data collection while reducing the logistical and financial burden typically associated with site-based monitoring.

Participants, identified in this study as Respondent 1 through Respondent 11 to preserve anonymity, were rural community members with limited ICT literacy and intermittent internet connectivity. All participants used their own smartphones to access and complete the form independently, without technical assistance. The design of the form emphasized simplicity, using straightforward language, checkboxes, and short-answer fields to accommodate users' familiarity with digital tools. Despite the technological and infrastructural challenges, participants successfully submitted multiple entries over a period of time, reporting key variables such as the number of chickens, types and timing of feed, signs of illness, and feed prices.

This repeated submission model enabled the researchers to gather longitudinal data for each participant, creating a continuous flow of information about poultry health and management practices. It also provided an opportunity to examine the consistency and completeness of data collected under realistic rural constraints. The methodology thus reflects a practical, low-cost strategy for digital monitoring in remote settings, offering valuable insights for future initiatives aiming to balance field engagement with the efficiencies of remote data collection.

4. Data Analysis and Data Quality Assessment

To evaluate the integrity and reliability of the dataset collected through the Google Form, a data quality assessment was conducted prior to any substantive analysis. The goal was to identify anomalies, inconsistencies, and patterns that might compromise the validity of the findings or hinder meaningful interpretation. Given that the form was designed primarily for monitoring purposes and completed independently by low ICT-literate users, special attention was paid to how usability, interface design, and user behavior influenced data quality.

The analysis process involved cleaning and reviewing 198 entries submitted by 11 unique respondents over multiple days. Microsoft Excel was used for initial data inspection, where sorting, filtering, and conditional formatting were employed to detect outliers and errors. The data was reviewed across four key quality dimensions: accuracy, completeness, consistency, and uniqueness [5]. These criteria are well-established in data management literature and are critical in ensuring that collected information can support both immediate monitoring and future research use [7].

Additionally, the characteristics of the digital data collection method were considered in interpreting anomalies. Digital forms filled out in remote and unsupervised settings, especially on smartphones, are susceptible to issues such as typing errors, autofill artifacts, and rapid or accidental submissions [1,3]. Given the participants' limited exposure to formal survey formats and data systems, deviations from structured input were anticipated and factored into the evaluation strategy.

5. Findings

The analysis of the dataset revealed several prominent data quality issues that must be addressed for both current interpretation and future data collection improvements.

a) Inconsistent Data Entry

One of the most significant challenges was the presence of mixed data types in fields intended for numerical input. For example, in the "Chicken Food Price" field, respondents entered a mix of numeric values (e.g., 12, 16.8) and textual entries (e.g., "Provided"). These inconsistencies complicated data cleaning procedures and impaired the dataset's suitability for quantitative analysis.

b) Duplicate and Rapid Submissions

Certain participants, particularly Respondent 1, submitted multiple entries within seconds of each other. These rapid-fire submissions raise questions about whether entries represented legitimate updates, accidental resubmissions, or technical issues such as repeated button presses. In the absence of unique identifiers or timestamps beyond submission time, distinguishing among these possibilities proved difficult.

c) Lack of Standardization in Categorical Fields

Several categorical fields, such as "Condition/Type of Illness," lacked predefined options, resulting in highly unstructured responses. While many respondents entered "None," others used free-text inputs like "kaki patah" or "monyok". Although these entries provide rich contextual insight, the lack of standard categories made aggregation and trend analysis more complex and less reliable.

d) Placeholder Values and Misinterpretation

Fields such as "Chicken Food Purchase" frequently contained entries such as "0" or "None." These may reflect the absence of a transaction on that day, but they could also be placeholders for skipped questions or misunderstood prompts. Without clarification or validation rules in the form, these ambiguous values reduce the overall interpretability of the dataset.

e) Absence of Unique Identifiers

The dataset used respondent names as the primary means of identification. For ethical reasons and to protect privacy, names were anonymized in this article as Respondent 1 through Respondent 11. However, the original absence of a unique respondent ID created difficulties in tracing patterns across time, especially where name variants or typographical inconsistencies existed. This limitation hindered accurate linkage of multiple submissions to the correct individual and impeded the ability to monitor behavioral trends.

6. Recommendation for Improving Data Quality

Based on the issues identified during analysis, several practical recommendations can be proposed to enhance data quality in future digital monitoring projects involving rural, low ICT-literate populations:

a) Introduce Data Validation and Input Restrictions

Implementing form-level constraints such as numeric-only input for price fields and dropdown menus for categorical options can prevent inconsistent data entry. This reduces the likelihood of mixed data types and enhances the dataset's readiness for quantitative analysis (Taherdoost, 2021).

b) Include Unique Identifiers and Session Metadata

Assigning each respondent a unique ID at the outset and automatically appending timestamp metadata can help differentiate between valid submissions and duplicates. This also improves the traceability and temporal analysis of responses (Hassenstein & Vanella, 2022).

c) Standardize Open-ended Fields

Where open responses are necessary, a limited list of pre-coded options with an "Other" field can balance flexibility and standardization. For example, illness types could be selected from a predefined list informed by common local poultry health issues, thereby enhancing comparability (Singh & Burgess, 2007).

d) Enhance User Training and Support Materials

Though participants completed the forms independently, providing short video instructions or visual guides through WhatsApp could help clarify common misinterpretations and ensure consistent form use across all respondents (Fitzgerald & FitzGibbon, 2014).

e) Periodic Data Verification

Occasional phone follow-ups or scheduled physical visits for verification of randomly selected entries can help ensure data reliability and build participant accountability. This hybrid approach is especially valuable where automated tools are insufficient.

7. Discussion

The findings underscore the complexity of collecting high-quality data in rural environments using digital self-report tools. While smartphone-based Google Forms provide a low-cost and scalable solution, the nature of the participants—low ICT-literate individuals with limited infrastructure—presents inherent challenges. The study reveals that even with minimal technological support, community members can participate in digital monitoring. However, the resulting data is vulnerable to common quality issues such as inconsistency, ambiguity, and duplication.

This case reflects broader concerns raised in digital research within low-resource contexts. As noted by Fitzgerald and FitzGibbon (2014), digital tools can foster participation and empowerment, but also risk introducing new layers of misunderstanding when not tailored to user capacity. Moreover, data quality management—often an overlooked step in digital fieldwork—becomes critically important in such contexts, particularly when data feeds directly into operational decision-making or research.

In this study, the primary value lies not only in the data collected but in the procedural insights into digital monitoring design. By surfacing quality issues and identifying root causes, this research contributes to the growing body of knowledge on electronic data collection in underserved regions and offers a pragmatic foundation for methodological improvement.

8. Conclusion

This study examined the use of smartphone-based Google Forms for daily remote monitoring of chicken farming activities among rural, low ICT-literate respondents. While the data collection strategy successfully enabled reduced site visits and empowered self-reporting, analysis revealed significant data quality issues, including inconsistent input formats, lack of standardization, ambiguous values, and absence of respondent identifiers.

Addressing these concerns requires a combination of technical design enhancements, respondent training, and hybrid verification strategies. The study offers practical lessons for researchers and practitioners aiming to implement digital data collection in similar contexts, emphasizing that the simplicity of tools must be matched with careful planning around usability, validation, and participant support.

Ultimately, the experience highlights the promise and pitfalls of digital surveys in rural development and research. When well-designed, digital tools can foster autonomy, reduce operational burdens, and provide rich longitudinal data. However, to realize these benefits, attention must be paid not just to access—but to design, support, and the inherent limitations of self-reported data under constrained conditions.

References

[1] M. Singh and S. Burgess, "Electronic data collection methods," in *Handbook of Research on Electronic Surveys and Measurements*, R. Reynolds, R. Woods, and J. Baker, Eds. Hershey, PA: IGI Global, 2007, pp. 28–43. doi: 10.4018/978-1-59140-792-8.ch004.

[2] H. Taherdoost, "Data collection methods and tools for research: A step-by-step guide to choose data collection technique for academic and business research projects," *Int. J. Acad. Res. Manage. (IJARM)*, vol. 10, no. 1, pp. 10–38, 2021.

[3] G. Fitzgerald and M. FitzGibbon, "A comparative analysis of traditional and digital data collection methods in social research in LDCs: Case studies exploring implications for participation, empowerment, and (mis)understanding," in *Proc. 19th IFAC World Congr.*, Cape Town, South Africa, Aug. 24–29, 2014, pp. 11437–11443.

[4] P. E. Choo, F. Mat Yamin, and W. H. Wan Ishak, "Customer satisfaction towards onsite restaurant interactive self-service technology (ORISST)," *Data Sci. Insights*, vol. 2, no. 1, pp. 35–44, 2024. [Online]. Available: https://citedness.com/index.php/jdsi/article/view/15

[5] M. J. Hassenstein and P. Vanella, "Data quality—Concepts and problems," *Encyclopedia*, vol. 2, no. 1, pp. 498–510, 2022. doi: 10.3390/encyclopedia2010032.

[6] M. F. M. Mohsin, W. H. Wan Ishak, Y. Yusof, J. Mohd Jamil, and A. Ahmad, "Web server log data pre-processing for mining zakat user profile using association rules," *Int. J. Bus. Intell. Data Min.*, vol. 26, no. 1/2, pp. 46–65, 2025. doi: 10.1504/IJBIDM.2025.143925.

[7] D. B. Shanmugam, J. Dhilipan, A. Vignesh, and T. Prabhu, "Challenges in data quality and complexity of managing data quality assessment in big data," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 9, no. 3, pp. 589–593, 2020. doi: 10.35940/ijrte.C5643.099320.

[8] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, no. 0, p. 2, 2015. doi: 10.5334/dsj-2015-002.

[9] M. I. Jaya, F. Sidi, L. S. Affendey, I. Ishak, and M. A. Jabar, "Systematic review of data quality research," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 21, 2019. doi: 10.5281/zenodo.5374485.

[10] F. Dravis, "Data quality strategy: A step-by-step approach (Practice-oriented paper)," in *Proc. 9th Int. Conf. Inf. Qual. (ICIQ-04)*, 2004, pp. 27–43.

[11] X. Li, L. Zhang, P. Zhang, and Y. Shi, "Problems and systematic solutions in data quality," *Int. J. Serv. Sci.*, vol. 2, no. 1, pp. 53–69, 2009.