International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(Research Article)

A web-based digital assistant for sample size determination and stratified sampling design

Farid Wahyudi *, Fajar Supanto and Harianto Respati

Doctoral program in economics, Merdeka University of Malang, Indonesia.

International Journal of Science and Research Archive, 2026, 20(01), 246–254

Publication history: Received on 30 May 2026; revised on 06 July 2026; accepted on 08 July 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.20.1.1454

Abstract

An adequate sample size and an appropriate sampling technique are frequently mischosen by students and empirical researchers. This study reports the development and internal validation of the Research Sampling Calculator & Assistant, a web-based decision support application for sample size determination and stratified sampling design. The application was developed as a client-side single-page web application (React, TypeScript, Tailwind CSS, and Vite) that implemented five sample size formulas (Slovin; Lemeshow; Cochran with and without finite population correction; Krejcie and Morgan; and the Hair et al. indicator-based multivariate ratio), a rule-based sampling-technique decision engine, a proportional and disproportional stratified allocation algorithm with an automated minority-stratum safeguard, and a local Excel-based random sampling engine that performs all computation inside the user's browser without transmitting participant data to a server. Formula outputs were validated against published reference tables and independent manual calculations. The disproportional allocation algorithm was tested on a simulated three-stratum population. Outputs from all five formulas were in perfect agreement with reference values. The disproportional module was able to eliminate minority-stratum underrepresentation while maintaining overall target sample size. Structured feature comparison with G*Power, Raosoft, SurveyMonkey, and OpenEpi shows that the developed application is the only one among the five that combines multi-formula guidance, interactive sampling-technique decision engine, auditable stratified allocation, and privacy-preserving local sample execution in a single freely accessible interface. The application gives a practical, validated, and privacy-conscious contribution to diminish common errors in sample size and sampling design in empirical research.

Keywords: Sample Size Determination; Stratified Sampling; Decision Support System; Research Methodology; Digital Assistant

1. Introduction

Sample size and sampling design are critical to the statistical power, generalizability, and validity of empirical research [1, 2]. A sample size that is too small increases the risk of inconclusive or unstable results, while a sample size that is too large leads to wasted resources and raises unnecessary ethical concerns, especially in health research [3, 4]. Sample size determination is one of the most commonly mishandled aspects of empirical research design, despite the existence of well-established formulae. Researchers often apply the Slovin formula to populations for which it was not intended, miscalculate sample size for infinite or unknown populations, and struggle to translate an overall target sample size into a defensible stratified allocation plan when minority subgroups are present in the population [5, 6]. Recent methodological audits have shown that a large proportion of published studies fail to provide adequate sample size justification or do not provide it at all [7, 8].

* Corresponding author: Farid Wahyudi

General-purpose tools solve parts of this problem, but not the whole thing. G*Power provides strict power-analysis-based sample size calculations but demands some statistical literacy (knowledge about the planned test, expected effect size, and power-analysis parameters) that many applied researchers have not yet acquired [9, 10]. Widely used web calculators such as Raosoft, the SurveyMonkey sample size calculator, and OpenEpi lower this barrier by exposing a single population-based formula, but none of these guide formula selection across differing population types, support indicator-based multivariate sample size rules used in structural equation modeling [11], or provide any mechanism for stratified allocation or for drawing an actual sample from a respondent list without uploading participant data to a remote server.

We describe the design, development, and internal validation of the Research Sampling Calculator & Assistant, a web-based application created to address this gap. This study had the following specific aims: (1) to implement and validate five well-established sample size formulas in a single, formula-selection-guided user interface; (2) to operationalize a rule-based sampling technique decision engine with an auditable disproportional stratified allocation algorithm that protects minority strata from underrepresentation; and (3) to implement this functionality in a privacy-preserving, fully client-side architecture where participant data are never transmitted to a server as prescribed by data-minimization principles emphasized in the software privacy engineering literature [12, 13].

2. Materials and methods

This study did not involve human participants or patient data or animals. All validation presented in Section 4 was performed using published reference tables, independently hand-calculated benchmark values, and a simulated (non-human) stratum population; thus, ethical committee approval was not required.

2.1. Architecture of the system

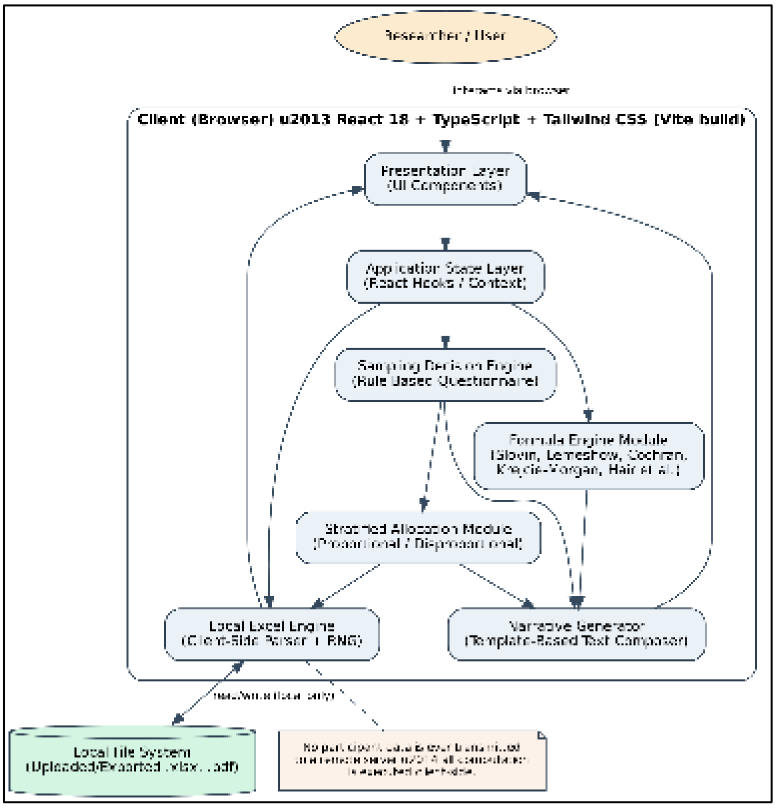

Figure 1 System architecture - Research Sampling Assistant & Calculator**.**

The application is a layered client-side single-page application (SPA) architecture (Figure 1) implemented in React 18, TypeScript, Tailwind CSS, and Vite. A Presentation Layer renders the user interface. An Application State Layer coordinates the flow of data among four independently testable modules: the Formula Engine, the Sampling Decision Engine, the Stratified Allocation module, and the Local Excel Engine. All four feed a Narrative Generator which composes an exportable methodology narrative. The only interaction with the local file system is a direct (browser-mediated) read/write for uploading a respondent list and exporting results. Participant data is never sent over the network.

2.2. Process flow

Figure 2 summarizes the control-flow logic from population-type selection to methodology narrative export, including the two governing decisions (finite/infinite/unknown population and stratified sampling) and the embedded minimum-viable-quota assessment that determines whether proportional or disproportional allocation is applied.

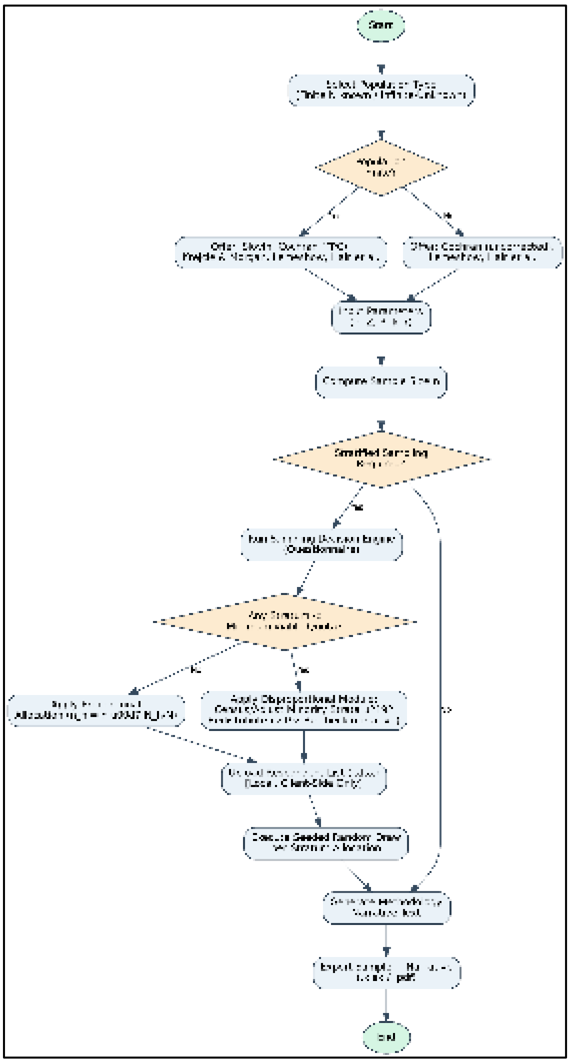

Figure 2 Methodology narrative export process flow. Population-type selection**.**

2.3. Class Organization

Figure 3 shows a simplified UML class diagram of the core domain model [14]. The PopulationInput class supplies validated parameters to the FormulaEngine. The SamplingDecisionEngine calls the StratifiedAllocation class, which has a set of Stratum objects. The LocalExcelEngine reads the allocation that was calculated in order to parse an uploaded respondent list and to perform a seeded random draw. The NarrativeGenerator combines the output from all of the previous classes.

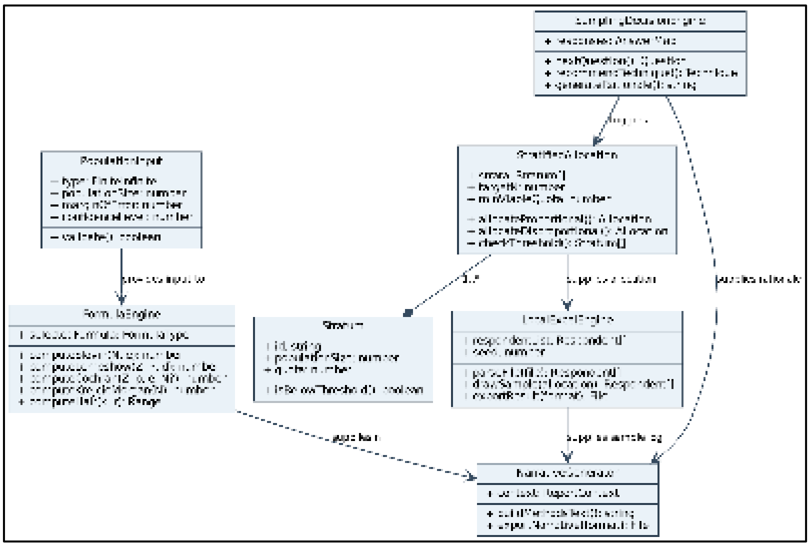

Figure 3 Simplified UML class diagram of the core domain model.

2.4. Sample size formulas

Five formulas were implemented as independent calculation modules, according to the standard presentation in the methodological literature.

2.5. Slovin's formula

Slovin's formula calculates the minimum sample size for a finite population with a given tolerable margin of error. It is only given for finite, defined populations:

$$n = N / (1 + N \cdot e^2)$$

2.6. Lemeshow's formula

Lemeshow’s formula calculates sample size on the basis of the expected proportion, confidence level and margin of error and can be applied to a finite or infinite population [15]:

$$n = Z^2 \cdot P(1 - P) / d^2$$

2.7. Cochran's formula

For large/infinite populations (without correction) Cochran's formula is used and for finite populations is automatically extended with the finite population correction:

$$n_0 = Z^2 \cdot p(1-p)/e^2 \qquad n = n_0 / (1 + (n_0-1)/N)$$

2.8. Krejcie & Morgan's formula

Krejcie and Morgan (1970) developed a formula for estimating the sample size from the chi-square distribution for a given confidence level which was cross-checked with the original 1970 lookup table:

$$n = \chi^2 NP(1-P) / [e^2(N-1) + \chi^2 P(1-P)]$$

2.9. Hair et al.'s multivariate ratio

To address the methodological criticism of rule-of-thumb PLS-SEM sample size heuristics [16,17,18], Hair et al. use an indicator-based multivariate ratio and report a recommended range based on the number of observed indicators, instead of a single number:

$$n = k \times r, \; r \in [5, 20]$$

2.10. Sampling-technique decision engine and stratified allocation

Figure 4 depicts the rule-based decision tree that was used to recommend a probability (simple random, systematic, stratified, cluster) or non-probability (purposive, quota, snowball, convenience) sampling technique from a short branching questionnaire, with a citable rationale statement [19,20]. Where stratified sampling is used, proportional allocation gives each stratum a quota proportional to its share of the population:

$$n_h = n \times (N_h / N)$$

A purely proportional rule gives the numerically minor strata a quota too small to be used. The disproportional module thus uses an iterative algorithm: (1) compare each stratum's proportional share with a user-defined minimum-viable-quota threshold; (2) assign a full census or an adjusted minimum quota to any stratum below threshold; (3) distribute the remaining sample budget proportionally among the remaining strata; and (4) repeat steps 1–3 until all strata are above the threshold or already fully censused.

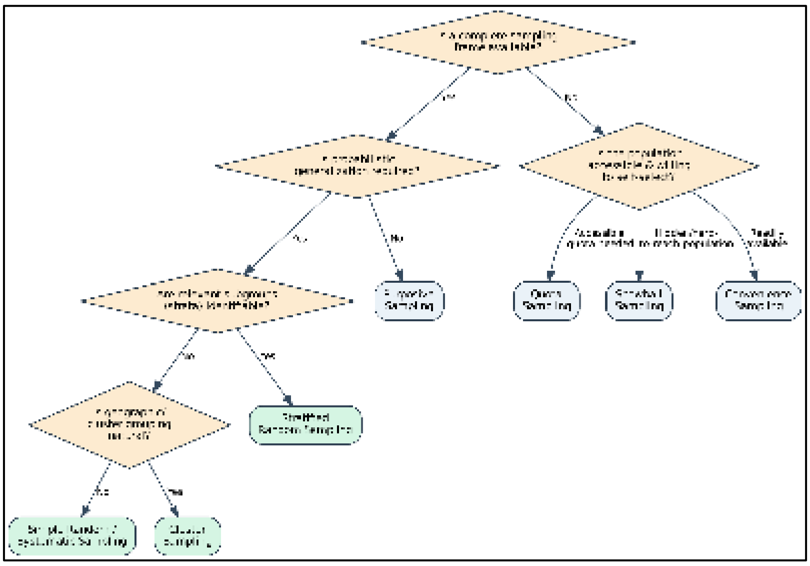

Figure 4 Decision tree implemented in the sampling-technique decision engine.

Algorithm (pseudocode)

```
FUNCTION allocateDisproportional(strata[], targetN, minQuota):
    active <- strata;  remainingN <- targetN;  result <- {}
    REPEAT
        converged <- TRUE
        totalPop <- SUM(s.populationSize FOR s IN active)
        FOR EACH s IN active:
            share <- remainingN * (s.populationSize / totalPop)
            IF share < minQuota:
                quota <- MIN(s.populationSize, minQuota)
                result[s.id] <- quota
                remainingN <- remainingN - quota
                REMOVE s FROM active;  converged <- FALSE
    UNTIL converged == TRUE
    totalPop <- SUM(s.populationSize FOR s IN active)
    FOR EACH s IN active:
        result[s.id] <- ROUND(remainingN * (s.populationSize / totalPop))
    RETURN result
END FUNCTION
```

2.11. Privacy-preserving local execution

The Local Excel Engine parses uploaded participant lists entirely client-side, and does random sampling with a seeded pseudo-random number generator to support reproducibility. No file content or respondent identifier is persisted on or transmitted to a remote server. This directly operationalises data-minimization principles that are increasingly emphasised in the privacy engineering literature [12, 13].

2.12. Validation procedure

The output of each formula module was validated using published reference values (Krejcie and Morgan 1970 table) and independently hand-calculated values within a matrix of population sizes (N = 100, 500, 1,000, 10,000, and infinite/unknown) and margins of error (e = 0.01, 0.05, 0.10). Validation was defined as exact numerical agreement. The disproportional allocation module was tested on a simulated three-stratum population (majority stratum 82%; minority strata 12% and 6%) with a target sample of n = 200 and a minimum-viable-quota threshold of 20, checking that (a) proportional allocations summed to n and (b) disproportional allocations eliminated any below-threshold stratum while preserving n.

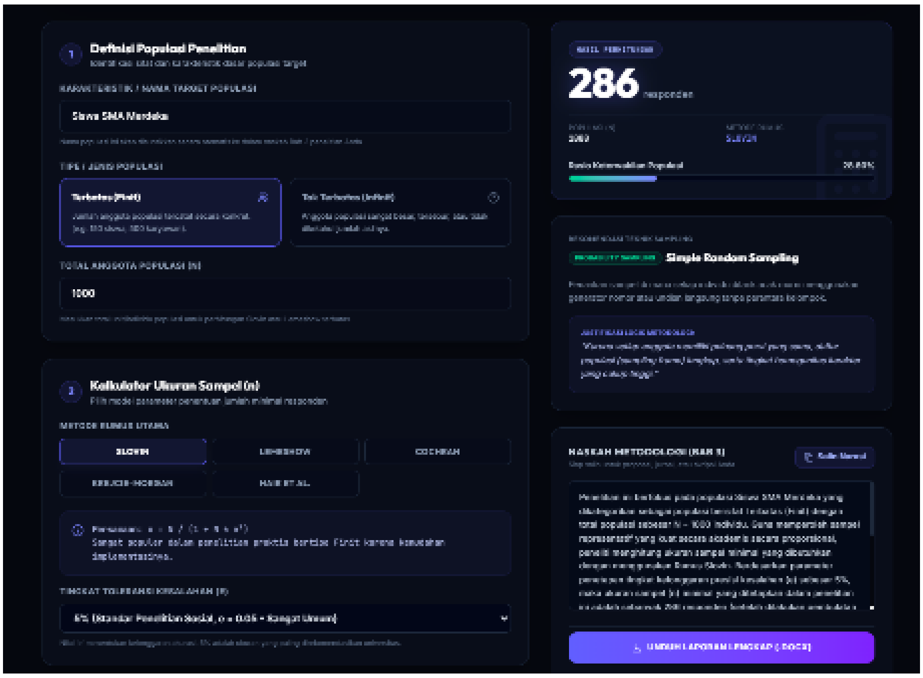

Figure 5 Population type selection and sample size formula input interface.

Figure 6 Proportional and disproportional stratified allocation table.

3. Results and discussion

Table 1 summarizes the comparative feature positioning of the developed application against G*Power, Raosoft, SurveyMonkey, and OpenEpi, based on their documented feature sets [9,10].

Table 1 Comparative feature positioning against G*Power, Raosoft, SurveyMonkey, and OpenEpi

Feature	G*Power	Raosoft	SurveyMonkey	OpenEpi	Developed app
Multiple sample size formulas	No (power-based)	Partial (1 formula)	Partial (1 formula)	Partial (epi-specific)	Yes (5 formulas)
Multivariate indicator ratio	No	No	No	No	Yes
Sampling-technique decision engine	No	No	No	No	Yes
Disproportional stratified allocation	No	No	No	No	Yes
Local, privacy-safe sampling	No	No	No	No	Yes
Methodology narrative export	No	No	No	No	Yes

The five formula modules were validated against published reference tables and manual calculation and, in every case tested, gave identical numerical results. Krejcie and Morgan outputs agreed with the original 1970 table for all population sizes checked (e.g., n = 384 at N ≥ 100,000, n = 278 at N = 1,000, and n = 80 at N = 100). The results of Slovin, Lemeshow, and Cochran matched the independently hand-calculated values for all combinations of population size and margin of error tested, and the Cochran finite population correction consistently reduced the uncorrected estimate n_0 in the expected direction as N decreased relative to n_0.

In the simulated three-stratum test, a target sample of n = 200 yielded 164, 24, and 12 respondents in the majority, first minority, and second minority strata, respectively, with the minimum-viable-quota threshold of 20; the smallest stratum (6% of the population) was below the threshold. The disproportional module increased this stratum to the 20-respondent minimum and allocated the remaining budget proportionally between the two active strata, satisfying the threshold condition for each stratum and maintaining the overall target sample size of n = 200.

These results show that the application developed is not an approximation but a digital operationalization of the well-established sampling theory. This is in line with the focus on rigorous and auditable sample size calculation in the literature reviewed [9, 3]. As shown in Table 1, the tool does not replace G*Power for formal power-analysis-based justification tied to a specific statistical test and effect size [9]. Yet, it meaningfully expands beyond Raosoft, SurveyMonkey, and OpenEpi by adding multivariate guidance, sampling-technique decision support, disproportional stratified allocation, and privacy-preserving local execution, addressing gaps also noted in adjacent software-engineering sampling literature [21, 22].

The disproportional stratified allocation results support the practical value of formalizing an otherwise ad hoc researcher judgment call [6] into a reproducible, auditable procedure, which may reduce the incidence of incomplete or absent sample size and sampling-design justification documented in recent methodological audits [23, 7, 24, 25, 26, 27]. The client-side, privacy-by-design architecture also makes the tool more attractive for research involving sensitive populations, since the raw participant lists never leave the user's device.

The present study has several limitations. Validation was limited to numerical agreement against reference values, not a field trial with practicing researchers, so usability and real-world error-reduction outcomes are still to be empirically established. The disproportional allocation test was conducted under one simulated three-stratum scenario. Further testing is needed to determine the behavior under a larger number of strata or more extreme minority proportions. As with all rule-of-thumb PLS-SEM heuristics, the Hair et al. multivariate ratio is best used as a starting range rather than as a replacement for a full power analysis where model complexity and expected effect sizes are known.

4. Conclusion

The study designed and internally validated a browser-based digital assistant that includes five validated sample-size formulae, an interactive sampling-technique decision engine, an auditable disproportional stratified allocation algorithm, and a privacy-preserving local sampling mechanism. All implemented formulas produced outputs in exact agreement with published reference values. The disproportional allocation module was successful in avoiding the under-representation of minority strata, in line with the target sample size. Comparison with four most popular tools shows the application meets an unmet need, combining multi-formula guidance, sampling-technique decision support, auditable stratified allocation, and privacy-preserving local sample execution all in a single open-access interface.

Compliance with ethical standards

Acknowledgments

The authors declare that there are no conflicts of interest. The authors would like to thank Merdeka University and Raden Rahmat Islamic University for providing the computing resources that were used in the development and testing of the application.

Disclosure of conflict of interest

The authors declare no financial or non-financial conflicts of interest in relation to this work. Individual disclosures: Farid Wahyudi – no conflict of interest. Fajar Supanto – no conflict of interest. Harianto Respati – no conflict of interest.

Statement of ethical approval

The present research work does not contain any studies performed on animals or human subjects by any of the authors.

Statement of informed consent

This study did not involve individual human participants; all validation was performed using published reference tables, hand-calculated benchmark values, and a simulated stratum population. Informed consent was therefore not applicable.

References

[1] Cochran WG. Sampling techniques. 3rd ed. New York: John Wiley & Sons; 1977.

[2] Krejcie RV, Morgan DW. Determining sample size for research activities. Educ Psychol Meas. 1970;30(3):607-10.

[3] Althubaiti A. Sample size determination: a practical guide for health researchers. J Gen Fam Med. 2023;24(2):72-8.

[4] Wang X, Ji X. Sample size estimation in clinical research. Chest. 2020;158(1):S12-20.

[5] Rahman MM. Sample size determination for survey research and non-probability sampling techniques: a review and set of recommendations. J Entrep Bus Econ. 2023;11(1):42-62.

[6] Andrade C. The inconvenient truth about convenience and purposive samples. Indian J Psychol Med. 2021;43(1):86-8.

[7] Nombera-Aznaran N, Guevara-Lazo D, Fernandez-Guzman D, Taype-Rondan A. Statistical characteristics of analytical studies published in Peruvian medical journals from 2021 to 2022: a methodological study. PLoS One. 2024;19(7):e0306334.

[8] Lakens D. Sample size justification. Collabra Psychol. 2022;8(1):33267.

[9] Kang H. Sample size determination and power analysis using the G*Power software. J Educ Eval Health Prof. 2021;18:17.

[10] Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175-91.

[11] Hair J, Alamer A. Partial least squares structural equation modeling (PLS-SEM) in second language and education research: guidelines using an applied example. Res Methods Appl Linguist. 2022;1(3):100027.

[12] Daoudagh S, Marchetti E, Savarino V, Bernal Bernabe J, Garcia-Rodriguez J, Torres Moreno R, et al. Data protection by design in the context of smart cities: a consent and access control proposal. Sensors. 2021;21(21):7154.

[13] De Chaves SA, Benitti FBV. Privacy by design in software engineering: an update of a systematic mapping study. Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing; 2023. p. 1362-9.

[14] Chen F, Zhang L, Lian X, Niu N. Automatically recognizing the semantic elements from UML class diagram images. J Syst Softw. 2022;193:111431.

[15] Lemeshow S, Hosmer DW, Klar J, Lwanga SK. Adequacy of sample size in health studies. Chichester: John Wiley & Sons; 1990.

[16] Kock N, Hadaya P. Minimum sample size estimation in PLS-SEM: the inverse square root and gamma-exponential methods. Inf Syst J. 2018;28(1):227-61.

[17] Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 8th ed. Andover: Cengage Learning; 2018.

[18] Hair JF, Hult GTM, Ringle CM, Sarstedt M. A primer on partial least squares structural equation modeling (PLS-SEM). 3rd ed. Thousand Oaks: Sage; 2022.

[19] Noor S, Tajik O, Golzar J. Simple random sampling. Int J Educ Lang Stud. 2022;1(2):78-82.

[20] Ting H, Memon MA, Thurasamy R, Cheah JH. Snowball sampling: a review and guidelines for survey research. Asian J Bus Res. 2025;15(1):1-15.

[21] Baltes S, Ralph P. Sampling in software engineering research: a critical review and guidelines. Empir Softw Eng. 2022;27(4):94.

[22] Russo D, Stol KJ. PLS-SEM for software engineering research: an introduction and survey. ACM Comput Surv. 2021;54(4):1-38.

[23] White M. Sample size in quantitative instrument validation studies: a systematic review of articles published in Scopus, 2021. Heliyon. 2022;8(12):e12223.

[24] Perdomo B, Morales OA. Errors and difficulties faced during the elaboration of undergraduate and graduate thesis by Peruvian students: pedagogical implications. Educare. 2022;26(1):380-400.

[25] Ahmed SK. How to choose a sampling technique and determine sample size for research: a simplified guide for researchers. Oral Oncol Rep. 2024;12:100662.

[26] Sadiq IZ, Usman A, Muhammad A, Ahmad KH. Sample size calculation in biomedical, clinical and biological sciences research. J Umm Al-Qura Univ Appl Sci. 2025;11:133-41.

[27] Zrineh A, Al-Usta M, Alwawi A. Sampling methods and sample size determination in clinical research: an educational review. J Gen Fam Med. 2026;27(1):e70096.