

Manual double entry of study data: potential to avoid 4 errors per 1000 entries

Kastytis Šmigelskas

Lithuanian University of Health Sciences

Monika Žemaitaitytė

`monika.zemaitaityte@lsmu.lt`

Lithuanian University of Health Sciences

Jūratė Koreivaitė

Lithuanian University of Health Sciences

Vladas Golambiauskas

Lithuanian University of Health Sciences

Research Article

Keywords: study data, double entry, error rate, error type, paper-to-digital data transfer

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10878982/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Previous research suggests that double entry is one of the most reliable methods for detecting entry errors. However, the effect of the usual single-entry method on errors is rarely assessed. The study aimed to examine data error rates detected during the manual double-entry process. An additional aim was to identify the main types of errors and their relation to human characteristics.

Methods

This study analyzed double-entry errors in 6,730 paper questionnaires from a large-scale study. Questionnaires were manually entered by 27 persons. We qualitatively identified and categorized error types. Error rates were calculated for each question and its characteristics. Non-parametric tests and robust regression were used to compare error rates and assess associations with question- and layout-related features.

Results

The error rate was 3.9 per 1,000 entries, with higher rates in open-ended questions. Response option size and matrix size were considered main predictors of error rate. In total, 29 error types were identified and divided into two main categories. Error rates were highest among international students and lowest among lab workers.

Conclusions

Double entry enables detecting and avoiding a small number of errors. The errors are more common in complex questions and relate to low experience of data entry clerks.

Background

Data accuracy is the cornerstone of reliable research. When data has to be transferred from paper to digital, a reliable entry method should be carried out to ensure its validity. As the amount of data increases with each additional participant queried, so does the number of errors. Therefore, picking the most reliable method for data accuracy is vital to minimize errors and their impact on the data.

Even though there are many methods to consider when ensuring the accuracy of the data collected, double entry is one of the most used methods whenever data must be transferred from one medium to another. It has also been proven to be relatively successful. For example, in clinical research databases, error rates that were detected by double entry ranged from 2.3 to 26.9%, indicating that the error rate can

get very high given the “correct” circumstances [1]. Other studies also find that, compared with single entry, double entry can ensure 8 [2] to 29 [3] times lower error rates. It was also found that the double-entry method is superior to the reading-aloud method or single entry with visual checking; the latter usually fails to detect most errors after single entry [3, 4, 5]. Other researchers still find that single entry with visual checking can potentially be even less accurate than just performing single entry [6]. Finally, Barchard & Pace [3] also discovered that only 0.06% of the errors discovered via double entry were blank or outside of the allowable range for the variables, suggesting most entry errors to be invisible after single entry. Overall, the evidence suggests that while many accuracy control strategies exist, double entry remains one of the best and widely used methods for accurate data transfers across mediums.

Research on data accuracy rarely considers or even acknowledges the types of errors made. However, there are mainly two categories to consider: interpretation errors and entry errors [1, 7]. Even though research on the error types is scarce, it suggests that interpretation errors are more common than entry errors [7]. In this study, open plain text fields that had to be interpreted by the entry staff had the highest error rate (1.2–2.1%), whereas open continuous variables and closed dichotomous variables had a significantly lower error rate (0.2–0.8% and 0.4–0.5%, respectively) [7]. However, the mentioned study can be considered outdated; thus, the error rate by error type remains unanswered.

The impact of entry errors cannot be overlooked. After all, if the impact is negligible, methods to assure data entry accuracy are pointless. A study by Barchard & Pace [3] suggests that even one single entry error, if big enough (e. g., changing 5 to 55), can have devastating effects when measuring correlations or differences between the means of independent groups. Even though detecting such values outside the range could be done without double entry, authors argue that it is still unclear how impactful within-range entry errors are (e. g., typing 4 instead of 1) [3]. In their own study, coefficient alpha differed by more than 0.10 from the correct value in at least one of the scales. Overall, the impact of entry errors depends on many factors but can be detrimental depending on circumstances and variables measured.

Even though prior studies highlight the value of the double-entry method, the cost of data entry and conflicting research should also be considered. Day, Fayers & Harvey [8] questioned whether the additional resources required for double entry are justified, showing that many detectable errors can instead be caught through range checks, logic tests, or simple exploratory data analysis. Gibson et al. in the CHART radiotherapy trials [9] found an error rate of 0.15% after single entry and only 0.025% for critical variables, concluding that the benefit of a second entry does not warrant the cost. Zhang & Hu [10] illustrated that improvement in single-entry accuracy greatly diminishes the error rate and the probability of detecting errors via double entry.

In terms of people-characteristic-dependent error rates, research by Czaja et. al. [11] indicates that age differences in data-entry and computer-based task performance are primarily reflected in speed rather than accuracy, with older adults typically working more slowly than younger adults and neither necessarily working with more or less accuracy than the other. However, the cited study is quite dated, and its conclusions may differ if repeated today.

Something worth considering when discussing data entry errors is the mechanisms behind them. Beyond mechanical or procedural errors, data entry accuracy can also be influenced by cognitive mechanisms. Repetitive manual tasks may affect the attention span and sometimes even cause cognitive “stickiness,” manifesting as perseveration – repeating previous responses even when the stimulus changes [12;13]. Experimental research shows that prolonged task exposure can increase perseveration during serial addition tasks [12], and that cognitive and motor perseveration are correlated, especially in older adults with reduced executive flexibility [14]. Even in healthy populations, perseveration occurs at measurable frequencies [15]. These findings suggest that certain recurring data-entry errors may stem directly from cognitive processes, rather than carelessness or mechanical errors.

Similarly, some complex errors resemble confabulation, where individuals unknowingly produce plausible but incorrect information. In healthy aging, confabulation has been linked to weak encoding and retrieval control – people fill gaps in memory with information that “feels right” [16]. In data entry, this could manifest when clerks reconstruct unreadable or missing responses based on expectation rather than the recorded value.

Thus, even though double entry seems the most reliable method to avoid data entry issues, the related costs can be quite high and depend on false entries avoided. The existing literature on the effect of double entry for data precision is very scarce; therefore, the aim of this study was to assess the error rate via double entry when transferring data from paper to digital. The secondary aim was to classify the most common types of errors, estimate their prevalence (frequency), and assess whether human characteristics play a part in making these errors.

Methods

This study analyzed the process of entering the paper questionnaires of the Health Behaviour in School-aged Children (HBSC) survey conducted in Lithuania in 2022 into the digital database. The entry process was performed in June-October 2022. In total, 6730 questionnaires were manually entered twice (by different people) into the “Microsoft Excel” database by 27 different data entry clerks (DECs). It took 5–10 minutes to enter one questionnaire into the database. Detailed data on the number of questionnaires entered by the DECs are presented in Table 1. Differences between entry 1 and entry 2 were identified as entry errors and were corrected through a third entry. During the verification process of data entry stages, any identified errors were documented using qualitative descriptions to identify the nature of the error, such as “rounded response”, “missing value”, “moved values”, etc.

Table 1
Characteristics of data entry clerks.

Local students	n	Shorter version of the questionnaire entries	Longer versions of questionnaire entries
	8	632	1381
International students	8	154	0
Lab workers	3	2747	3365
Researchers	8	2645	1984

During the survey, students were asked to fill out an anonymous paper questionnaire in A5 format, answering questions about their lifestyle, health and well-being, risky behavior, and social environment. The tool for the study was based on the then-current version of the standardized international HBSC research protocol [17]. Two versions of the questionnaire were used. 5th- and 7th-grade students received a shorter, 20-page questionnaire that consisted of 65 questions. In total, this version had 189 items, of which 4 were open-ended. 9th- and 11th-grade students received a longer, 24-page questionnaire that consisted of 80 questions. In this version of the questionnaire, students filled in 219 items, of which 5 were open-ended. The questionnaire included various types of questions: open-ended, closed-ended with various numbers of response options, scale-based questions, and matrix questions. Most closed-ended questions were coded from 1 at the time of entry, but there were a few questions that were coded from 0 or vice versa (e.g., from 5 to 1).

Data analysis

To simplify calculations, both the shorter and the longer questionnaires were merged into a single database. Errors per 1000 entries (error rate) were calculated for each question and its characteristic (e.g., number of response options, number of items per block, etc.). The number of missing data points in the questionnaire was considered when calculating the error rate, but the error rate was based exclusively on errors made within the completed data.

For the comparison of error rates between open-ended and closed-ended question types, the assumptions of parametricity were considered violated. This decision was based on the highly unequal question group sizes (399 vs. 9) and the unjustifiable normality assessment for the smaller group. Consequently, the non-parametric Mann–Whitney U test was chosen for data comparison.

When comparing error rates based on question characteristics, the assumption of normality was tested using the Kolmogorov–Smirnov test. Within these groups, the data distribution significantly deviated from normal ($p < .05$), leading to the selection of Spearman's correlation. To investigate specific relationships, Spearman's correlation was conducted to determine the relationship between the error rate and item characteristics such as response option type, matrix size, and item sequence within the

questionnaire. A robust linear regression analysis was performed using MM-estimation to ensure resistance to outliers. Two robust linear regression models were used to examine how question size, response option size, and item sequence predict error rates and how matrix size and item sequence predict error rates. To determine statistical significance, an asymptotic t-test with $n-k$ degrees of freedom was employed. The p-values were calculated using a two-tailed Student's t-distribution according to the formula: $p = 2 \cdot P(T_{df} > |t|)$. Additionally, a Kruskal-Wallis test was used to examine the relationship between item positioning on the page (left page/right page, upper/lower part) and the error rate.

Based on qualitative error descriptions, error types were identified and clustered into categories. For each category, average impact (average number of entries affected by entry error) and error rate were calculated.

Results

Error rate by question and its characteristic

Total error rate per 1000 entries was 3.9. Error rates by question varied from 0.1 to 25.4. When evaluating the number of errors by question type, error rates were significantly higher for open-ended (Mdn = 10.9, IQR = 7.4–17.2) than close-ended questions (Mdn = 3.3, IQR = 1.9–5.1), $U = 147.0$, $Z = -4.71$, $p < .001$. Examples of open-ended items with particularly high error rates included weight (7.7–25.4), height (8.4–20.0), and number of siblings (7.1–14.4).

First, the overall correlation between error rate and closed-ended question characteristics was assessed. A Spearman's correlation was run to determine the relationship of error rates with question size (number of items within one question), response options size (Fig. 1), and matrix size (items $\times$ responses) (An additional file shows this in more detail [see Additional file 1]). A weak positive correlation was found in relation to response options size ($r = .17$, $p < .001$). Other correlations were non-significant: error rate was not associated with question size ($r = .04$, $p = .423$), matrix size ($r = .09$, $p = .082$), and item sequence ($r = -.06$, $p = .242$).

A robust linear regression model (MM estimation) was used to examine how question size, response options size, and item sequence predict the error rate. Results show that the number of response options consistently emerged as the strongest and most reliable predictor of error rate, indicating that each additional answer option increases the expected error rate by 0.142, while keeping the other variables constant. In contrast, the effects of question size and item sequence in the questionnaire were not statistically significant (Table 2).

In the second robust regression model, in which item order and matrix format structure were considered as predictors of error rates, matrix size shows a statistically significant positive effect, while item sequence in the questionnaire does not appear significant (Table 2). Each additional matrix increment by

one value increases the expected error rate by .014, while keeping the item sequence in the questionnaire constant.

Table 2
Factors predicting error rate

		<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Model 1	Question size	0.042	0.025	1.678	0.094
	Response options size	0.142	0.056	2.525	0.012
	Item sequence	-0.001	0.002	-0.574	0.566
	Constant	2.624	0.417	6.292	< 0.001
Model 2	Matrix size	0.014	0.004	3.209	0.001
	Item sequence	-0.001	0.002	-0.770	0.442
	Constant	3.196	0.258	12.383	< 0.001

In Model 1, residual standard error: 2.032 on 395 degrees of freedom. In Model 2, residual standard error: 2.028 on 396 degrees of freedom.

The questionnaire was designed as an A5 format booklet printed on both sides. This means that the respondents, as well as the data clerks, were paging the questionnaire, which could lead to potential changes in their attention or other mental capacity. To check this, item positioning within double-page format (upper left, lower left, upper right, and lower right quadrants) was correlated with error rate. A Kruskal-Wallis test indicated no significant differences in error rates by item positioning by quadrant ($\chi^2(3, N = 399) = 1.274, p = .735$). The median error rate was 3.2 for the upper left, 3.7 for lower left, 3.3 for upper right, and 3.0 for lower right quadrant.

Error types

Based on the qualitative error descriptions, 29 error types were identified and divided into two main categories based on the origin of the errors that depend on data entry clerk decisions (omission, partial data entry, alteration, correction handling, and transcription errors) and errors related to cognitive-motor interference (typing, cognitive and memory, redundant or composite, positioning errors). Error types and descriptions are shown in Table 3.

Table 3
Types of errors

Error types			Description
Data entry clerk decisions	Omission Errors	Single Value Omission	The respondent indicated a single value response but it was not entered
		Multiple Values Omission	The respondent indicated several values, but no value was entered
		Partial Multiple Values Omission	The respondent provided multiple values in the question but one of the values was not entered
	Partial Data Entry Errors	First-Value Recording	The respondent indicated several values but only the first value was entered (e.g. in place of 123→1)
		Last-Value Recording	The respondent indicated several values but only the last one was entered (e.g. in place of 123→3)
	Alteration Errors	Response Alteration Error	In questions that adolescents could interpret differently, the facilitator immediately corrected them, rather than marking the original input (e.g. instead of the days of the week 1, 2, 5, the number of days entered is 3)
		Averaging Error	The respondent indicated several values but the average was entered instead (e.g. in place of 12→1.5)
		Rounding Error	The exact answer of the respondent was rounded (e.g. 175.8 → 176)
	Correction Handling Errors	Dual Value Entry	The respondent crossed out the value and marked the correct one but both the correct and crossed out values were entered
		Incorrect Value Entry	The respondent crossed out the value and marked the correct one but only the crossed-out value was entered
		Crossed-Out Value Entry	The respondent crossed out the value, but it was entered in the database nonetheless
	Transcription Errors		Handwritten digit or marking was misinterpreted when entering the data
Cognitive-motor interference	Mechanical /Typing Errors	Horizontal key press	Horizontally adjacent button on the numeric keypad bar was pressed (e.g. in place of 1→2, 2→3, 4→5)
		Vertical key press	Vertically adjacent button on the numeric keypad bar was pressed (e.g. in place of 1→4, 2→5, 3→6)

Error types		Description	
Cognitive and memory Errors	Diagonal key press	The diagonal button on the numeric keypad bar was pressed (e.g. in place of 1→5, 2→4, 2→6)	
	Numeric row key press	An adjacent button on the numeric row above the keyboard letters was pressed (e.g. in place of 3→4, 6→7, 9→0)	
	Invented coding	The respondent selected an answer that is semantically correct, but due to a misinterpretation of the scale (e.g., reverse order or starting from 0), an incorrect code was entered	
	Perseveration	Correct value from previous adjacent question was entered instead of the current question	
	Anticipation error	A correct value of the subsequent question was entered instead of the current question	
	Confabulation	A value was entered that did not exist at all and there is no way to explain it	
	Redundant or Composite Entry Errors	Duplicate Value Entry	Correct value was entered twice in the same cell
		Previous Value Addition	Correct value from previous adjacent question was entered in addition to the correct value from the current question
		Next Value Addition	A correct value of the subsequent question was entered in addition to the correct value from the current question
		Extraneous Value Entry	A false value not from an adjacent question was entered in addition to the correct value from the current question
Positioning Errors	Split value error	Two values of one question were entered in two separate cells (e.g. in place of 12 -> 1, 2)	
	Merged values error	The values of two adjacent questions were written in one cell, the adjacent cell was left blank (e.g. in place of 1, 2 -> 12, missing)	
	Transposition errors	The values of two adjacent answers were swapped with each other (e.g. in place of 1, 2 -> 2, 1)	
	Cascading error	One value was missed while entering, the subsequent entries followed the wrong series (e.g. in place of 1, 2, 3, 4, 5, 6 -> 2, 3, 4, 5, 6 missing)	
Other			

The analysis of data entry errors identified distinct patterns between the frequency of errors (errors per 1000 entries) and their severity (average impact to entry). The highest overall frequency of errors was attributed to cognitive-motor interference, which recorded an error rate of 2.7, more than twice as frequent as data entry clerk decision errors (1.11). Cognitive-motor interference errors were mainly caused by positioning errors (1.14), dominated by cascading errors, where one value was omitted during entry and subsequent entries were made in the wrong order (0.94), and mechanical/typing errors (0.87), dominated by horizontal keystrokes, where a horizontally adjacent key on a numeric keypad was pressed (0.55). Meanwhile, the data entry clerk decision category was dominated by omission errors (0.55), most commonly single value omission (0.44).

Although cognitive-motor interference occurred more frequently, its average impact on recordings (1.9) was only slightly greater than the impact caused by the data entry clerk's decisions (1.4). However, the two groups showed different trends within the subcategories. The greatest threat to data integrity was posed by positioning errors, with an average impact of 4.0 across the category. Notably, cascading errors had the highest individual impact score of 4.9. More comprehensive results regarding types of errors are shown in Table 4.

Table 4
Frequency and severity of different error types

Type of error	Average impact to entries	Error rate
1. Data entry clerk decisions	1.4	1.11
1.1. <i>Omission Errors</i>	1.9	0.55
1.1.1. Single Value Omission	2.0	0.44
1.1.2. Multiple Values Omission	1.6	0.08
1.1.3. Partial Multiple Values Omission	1.5	0.02
1.2. <i>Partial Data Entry Errors</i>	1.2	0.15
1.2.1. First-Value Recording	1.3	0.10
1.2.2. Last-Value Recording	1.1	0.05
1.3. <i>Alteration Errors</i>	1.1	0.12
1.3.1. Response Alteration Error	1.0	0.01
1.3.2. Averaging Error	1.1	0.11
1.3.3. Rounding Error	1.0	0.01
1.4. <i>Correction Handling Errors</i>	1.0	0.25
1.4.1. Dual Value Entry	1.0	0.13
1.4.2. Incorrect Value Entry	1.0	0.05
1.4.3. Crossed-Out Value Entry	1.2	0.06
1.5. <i>Transcription Errors</i>	1.0	0.04
2. Cognitive-motor interference	1.9	2.68
2.1. <i>Mechanical /Typing Errors</i>	1.4	0.87
2.1.1. Horizontal key press	1.4	0.55
2.1.2. Vertical key press	1.3	0.09
2.1.3. Diagonal key press	1.4	0.07
2.1.4. Numeric row key press	1.6	0.15
2.2. <i>Cognitive and memory Errors</i>	1.3	0.60
2.2.1. Invented coding	1.9	0.18
2.2.2. Perseveration	1.1	0.13
2.2.3. Anticipation error	1.1	0.15

Type of error	Average impact to entries	Error rate
2.2.4. Confabulation	1.2	0.13
2.3. <i>Redundant or Composite Entry Errors</i>	1.0	0.07
2.3.1. Duplicate Value Entry	1.0	0.02
2.3.2. Previous Value Addition	1.0	0.02
2.3.3. Next Value Addition	1.0	0.01
2.3.4. Extraneous Value Entry	1.0	0.03
2.4. <i>Positioning Errors</i>	4.0	1.14
2.4.1. Split value error	2.0	0.03
2.4.2. Merged values error	2.0	0.01
2.4.3. Transposition errors	2.1	0.16
2.4.4. Cascading error	4.9	0.94
Other	2.4	0.04

Data entry clerk (DEC)

Depending on the occupation of DEC, error rate per one HBSC questionnaire varied from 0.3 to 5.6. Lab workers had the lowest error rates at 0.3 errors per questionnaire, followed by researchers (1.0) and local students (1.1). Foreign students had the greatest error rate at 5.6 errors per questionnaire.

The analysis of error types across DEC groups reveals distinct patterns: positioning errors were overwhelmingly the single most frequent type for international students, accounting for nearly 68% of their errors. While local students made positioning errors twice as less (30.7%), they had a high frequency of mechanical (24.3%) and cognitive memory (19.6%) errors. Researchers mostly made positioning (26.0%) and mechanical errors (21.1%), but were distinguished by the highest frequency of omission errors (19.0%) and alteration errors (6.1%), while lab workers mostly made mechanical (31.1%) and positioning errors (22.4%), and had a significant number of cognitive and memory (19.6%) errors (Table 5).

Table 5
Frequency of error types be data entry clerks

Types of errors, %		Local students	International students	Lab workers	Researchers
Data entry clerk decisions	Omission Errors	10.8	8.9	8.6	19.0
	Partial Data Entry Errors	2.2	2.1	4.7	4.8
	Alteration Errors	0.5	0.9	0.2	6.1
	Correction Handling Errors	10.1	3.2	8.4	4.6
	Transcription Errors	1.2	0.2	1.7	0.9
Cognitive-motor interference	Mechanical / Typing Errors	24.3	7.3	31.1	21.1
	Cognitive and memory Errors	16.1	8.4	19.6	15.0
	Redundant or Composite Entry Errors	2.8	0.7	2.4	1.4
	Positioning Errors	30.7	67.9	22.4	26.0
Other		1.3	0.2	0.7	1.3
Total errors per one questionnaire		1.1	5.6	0.3	1.0

Discussion

Main findings

This study identified and quantified data entry errors and assessed how they relate to the characteristics of items, data entry clerks, and the data itself. It also assessed the impact that errors have on the validity of the data.

The overall error rate was 3.9 per 1000 entries, comparable to rates reported in large-scale clinical and epidemiological databases [18]. Cognitive-motor interference type errors were more than twice as frequent as data entry clerk decisions type errors. The most error-prone items were open-ended numeric responses such as height, weight, and number of siblings, aligning with previous evidence that numeric-entry and free-text fields yield higher error probabilities than categorical items [1, 2].

While the absolute error rate may seem small, even a single incorrect numeric value can meaningfully distort statistical estimates. Barchard & Pace [3] demonstrated that one large mis-entry can alter

correlation coefficients or group-mean differences by over 0.10. In our study, across 29 identified error types, mechanical slips (e. g., adjacent-key errors) and cascading sequence shifts (a case when one value is missed and the subsequent entries follow the wrong series) dominated, while higher-order cognitive errors such as perseveration or confabulation were less frequent. Cascading sequence shifts were the most impactful error type, impacting 5 subsequent entries.

Errors by characteristics

Correlation analysis revealed that question response size (number of responses given to the question) was weakly associated with error rate, while question size (number of questions per item), matrix size (number of questions x answers per item), and item sequence (order of the question in a questionnaire) were insignificant factors. Question response size emerged as the single significant factor (albeit a weak one) impacting the error rate in the robust regression models as well. This suggests that entries are more prone to error with an increasing number of responses in an item. The position analysis revealed no page-layout effect, suggesting that item placement within the page has no effect on the error rate.

Finally, data entry error rate varied significantly depending on the occupation of the data entry clerk (DEC). Lab workers, who are the most experienced with data entry, exhibited the lowest error rate (0.3), followed by researchers (1.0) and local students (1.1). Foreign students exhibited the greatest error rate of 5.8 errors per one questionnaire. These findings suggest that familiarity with the task reduces the error rate significantly while the language barrier is impactful even when dealing with mainly numerical entries. Structured training programs emphasizing numeric keypad ergonomics, verification routines, and familiarity with questionnaire logic could therefore yield measurable quality gains at low cost. Assuming that the error rate would drop to the level of the lab workers, double entry could even be considered obsolete.

When deciding the need for data entry error screening, it's important to consider the cost associated with each method. For double entry, the data entry task essentially doubles the time needed to perform it. Considering it takes 5–10 minutes to input one questionnaire, double entry cost this study at least 561 extra hours. Even though double entry was consistently found to be most reliable at error detection [2, 3], the error rate in our study was rather low and not very impactful. Therefore, when dealing with large databases, researchers should weigh the options they have for error screening and prevention, and consider the implications associated with it carefully before making their final decision.

Limitations

Several limitations should be considered. First, error detection relied on double entry comparison; therefore, systematic misinterpretations repeated by both clerks could have remained undetected. Second, these findings originate from one national HBSC dataset and may not generalize to other survey designs, languages, or data entry interfaces.

Conclusions

This study showed that using the double-entry method resulted in a low overall error rate, although errors varied considerably by question type and structure. Open-ended questions, larger numbers of response options, and larger matrix formats were associated with higher error rates. 2 main categories of error types were identified. Although cognitive-motor interference occurred more frequently, its impact was only slightly greater than the impact of data entry clerk decisions. Differences between data entry clerk occupations suggest that human-related factors may also contribute to error occurrence.

Declarations

Ethics approval and consent to participate

This study analyzed the process of entering the paper questionnaires of the Health Behaviour in School-aged Children (HBSC) survey conducted in Lithuania in 2022 into the digital database, but did not analyze the direct survey data. The HBSC study was conducted following guidelines from the Declaration of Helsinki and approved by the Bioethics Centre of the Lithuanian University of Health Sciences (protocol No. BEC-M-05, date of approval 17 March 2022). The researchers followed all ethical guidelines for information processing. All subjects involved in this study provided verbal informed consent, while a passive informed consent procedure was employed to inform students' parents about the HBSC study and provide them with the opportunity to refuse participation.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to the sensitivity of the data, which includes health and behavioural information collected from minors, and the requirement to protect participant confidentiality in accordance with ethical approval conditions, but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

Authors' contributions

K.Š.: conceptualization, methodology, supervision, writing—review, and editing. M.Ž.: conceptualization, investigation, methodology, formal analysis, writing—original draft. J.K.: conceptualization, methodology, investigation, formal analysis, writing the original draft, writing—review, and editing. V.G.: conceptualization, methodology, investigation, writing the original draft, writing—review, and editing. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

Not applicable.

References

1. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. *AMIA Annu Symp Proc*. 2008 Nov 6;2008:242-6. PMID: 18998889; PMCID: PMC2656002.
2. Paulsen A, Overgaard S, Lauritsen JM. Quality of data entry using single entry, double entry and automated forms processing—an example based on a study of patient-reported outcomes. *PLoS One* 2012;7:e35087. doi: 10.1371/journal.pone.0035087.
3. Barchard KA, Pace LA. Preventing human error: the impact of data entry methods on data accuracy and statistical results. *Comput Human Behav* 2011;27:1834-9. doi: 10.1016/j.chb.2011.04.004.
4. Kawado M, Hinotsu S, Matsuyama Y, Yamaguchi T, Hashimoto S, Ohashi Y. A comparison of error detection rates between the reading aloud method and the double data entry method. *Control Clin Trials* 2003;24:560-9. doi: 10.1016/S0197-2456(03)00089-8.
5. Barchard KA, Scott J, Weintraub D, Pace LA. Better data entry: double entry is superior to visual checking. Paper presented at: American Psychological Association Annual Convention; 2008 Aug 14-17; Boston, MA. Accessed 20 Feb 2026. Available from: unlv.edu.
6. Scott JR, Thompson AR, Wright-Thomas D, Xu X, Barchard KA. Data entry methods: is double entry the way to go? Poster presented at: Western Psychological Association Annual Convention; 2008 Apr 10-13; Irvine, CA. Accessed 20 Feb 2026. Available from: unlv.edu.
7. Büchele G, Och B, Bolte G, Weiland SK. Single vs. double data entry. *Epidemiology* 2005; 6:130-1. doi: 10.1097/01.ede.0000147166.24478.f4.
8. Day S, Fayers P, Harvey D. Double data entry: what value, what price? *Control Clin Trials* 1998;19:15-24. doi: 10.1016/S0197-2456(97)00096-2.
9. Gibson D, Harvey AJ, Everett V, Parmar MKB. Is double data entry necessary? The CHART trials. *Control Clin Trials* 1994;15:482-8. doi: 10.1016/0197-2456(94)90005-1.
10. Zhang J, Hu W. Single or double data entry: considerations based on a simple binomial model. *Control Clin Trials* 1998;19:56-8. doi: 10.1016/S0197-2456(97)00122-0.
11. Czaja SJ, Sharit J. Ability-performance relationships as a function of age and task experience for a data entry task. *J Exp Psychol Appl* 1998;4:332-51. doi: 10.1037/1076-898X.4.4.332.

12. Giannouli V. Number perseveration in healthy subjects: does prolonged stimulus exposure influence performance on a serial addition task? *Adv Cogn Psychol* 2013;9:15-9. doi: 10.2478/v10053-008-0127-8.
13. Tomei G, Cinti ME, Cerratti D, Fioravanti M. Attenzione, lavori ripetitivi, fatica e stress [Attention, repetitive works, fatigue and stress]. *Ann Ig* 2006;18:417-29. PMID: 17089957.
14. Sombric CJ, Torres-Oviedo G. Cognitive and motor perseveration are associated in older adults. *Front Aging Neurosci* 2021;13:610359. doi: 10.3389/fnagi.2021.610359.
15. Ramage A, Bayles K, Helm-Estabrooks N, Cruz R. Frequency of perseveration in normal subjects. *Brain Lang* 1999;66:329-40. doi: 10.1006/brln.1999.2032.
16. Attali E, Dalla Barba G. Confabulation in healthy aging is related to poor encoding and retrieval of over-learned information. *Aging Neuropsychol Cogn* 2013;20:339-55. doi: 10.1080/13825585.2012.711462.
17. Inchley J, Currie D, Samdal O, Jåstad A, Cosma A & Nic Gabhainn S, editors. Health Behaviour in School-aged Children (HBSC) Study Protocol: background, methodology and mandatory items for the 2021/22 survey. Glasgow: MRC/CSO Social and Public Health Sciences Unit, University of Glasgow; 2023.
18. Garza MY, Williams T, Ounpraseuth S, Hu Z, Lee J, Snowden J, et al. Error rates of data processing methods in clinical research: a systematic review and meta-analysis of manuscripts identified through PubMed. *Int J Med Inform* 2025;195:105749. doi: 10.1016/j.ijmedinf.2024.105749.

Figures

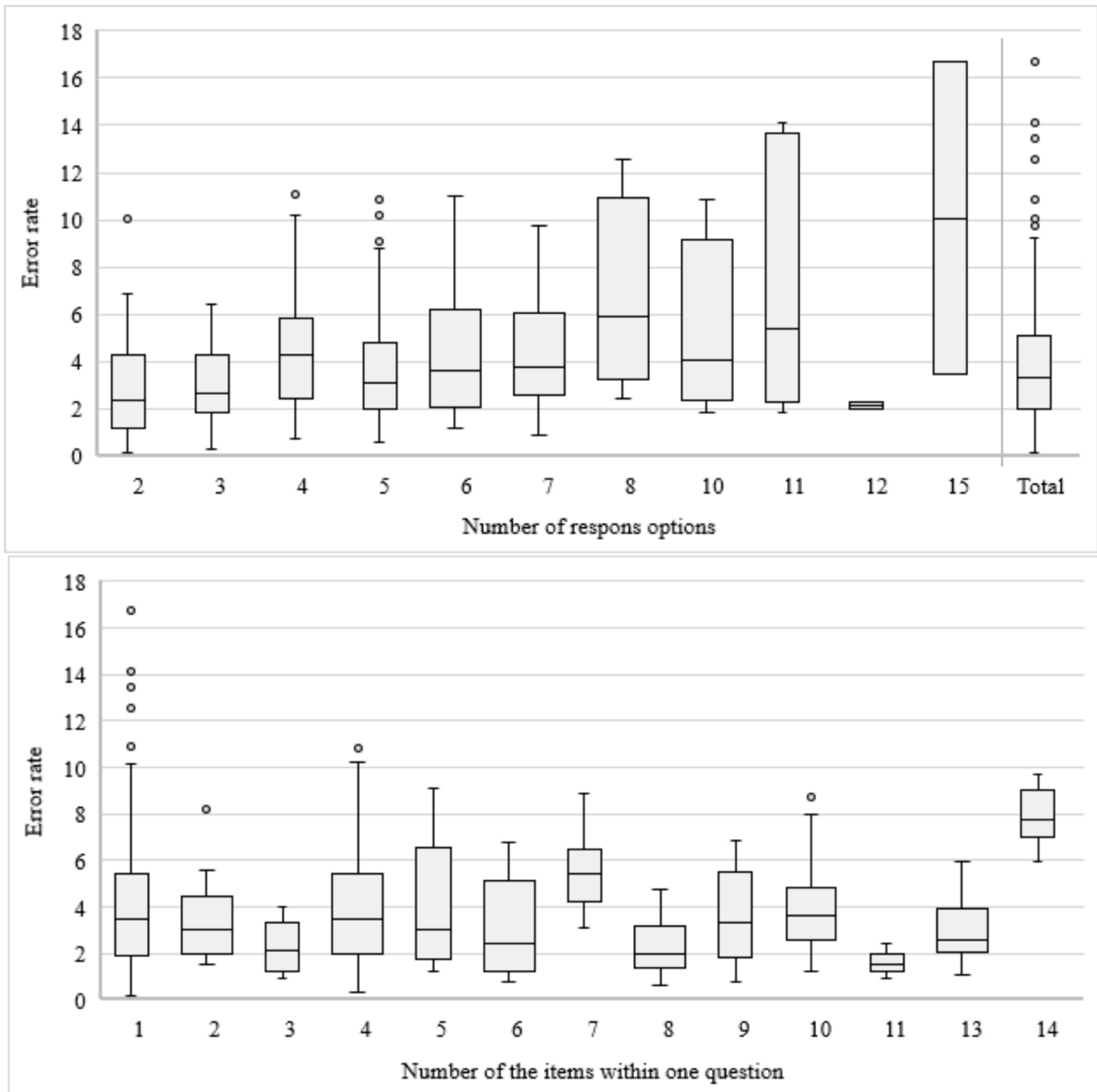


Figure 1

Correlations Between Error Rate and Question Size

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)