Students’ Semantic Mistakes in Writing Seven Different Types of SQL Queries

Alireza Ahadi, Julia Prior, Vahid Behbood and Raymond Lister

University of Technology, Sydney, Australia

{Alireza.Ahadi, Julia.Prior, Vahid.Behbood, Raymond.Lister}@uts.edu.au

ABSTRACT

Computer science researchers have studied extensively the mistakes of novice programmers. In comparison, little attention has been given to studying the mistakes of people who are novices at writing database queries. This paper represents the first large scale analysis of students’ semantic mistakes in writing different types of SQL SELECT statements. Over 160 thousand snapshots of SQL queries were collected from over 2300 students across nine years. We describe the most common semantic mistakes that these students made when writing different types of SQL statements, and suggest reasons behind those mistakes. We mapped the semantic mistakes we identified in our data to different semantic categories found in the literature. Our findings show that the majority of semantic mistakes are of the type “omission”. Most of these omissions happen in queries that require a JOIN, a subquery, or a GROUP BY operator. We conclude that it is important to explicitly teach students techniques for choosing the appropriate type of query when designing a SQL query.

General Terms

Performance, Human Factors.

Keywords

Online assessment; databases; SQL queries.

1. INTRODUCTION

The Structured Query Language (SQL) is the standard language for relational and object-oriental databases, as well as the industry standard language for querying databases. As with other computer languages, SQL queries can be semantically or syntactically wrong. However, limited attention has been given to understanding novice programmers’ challenges in writing correct SQL queries [5]. A deep understanding of the common semantic mistakes that novices make when writing SQL queries will improve teaching and learning outcomes.

ITiCSE '16, July 09-13, 2016, Arequipa, Peru

© 2016 ACM. ISBN 978-1-4503-4231-5/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2899415.2899464

In this paper, we use data collected over nine years from ~2300 students taking online SQL exams. One of the ways that we have analyzed this data is to qualitatively study the semantic mistakes committed by these students. We review these mistakes in seven different types of SQL queries and investigate the reasons behind them. More specifically, we map these mistakes to proposed mistake categories introduced in the literature, and explain why students are likely to make these mistakes.

In section 2, we review the literature on analysis of errors in SQL. In Section 3, we describe the data collection and analysis of the data to explore the types of SQL query errors made by students. In section 4, we review our findings on common semantic mistakes of novices in writing different SQL SELECT statements. Section 5 expands and discusses these findings, before our conclusions are given in Section 6.

2. RELATED WORK

Most computing education researchers who have studied database education have focused on tutoring and/or assessment tools. Most of that work has been concerned with the functionality of the tool itself, or on how the system supports a certain pedagogical model [1-4]. There has been relatively little work on novice errors and misconceptions when using SQL.

A few studies have investigated these challenges with programming in SQL. Reisner performed the first experimental study investigating SEQUEL, the predecessor of SQL [7]. In that study, a series of psychological experiments were conducted on college students to investigate learnability of the language, as well as the type and frequency of errors made by subjects. Reisner categorized students’ mistakes into “intrusion”, “omission”, “prior-knowledge”, “data-type”, “consistency” and “over-generalization”.

Welty and Stemple [8] explored users’ difficulty in writing queries in SQL compared to TABLET. Their comparison revealed that constructing difficult queries in more procedurally-oriented languages was easier than less procedurally-oriented languages. They categorized SQL statements into “correct”, “minor language error”, “minor operand error”, “minor substance error”, “correctable”, “major substance error”, “major language error”, “incomplete” and “unattempted”, where the first four of those categories were considered *essentially correct* and the other five categories were classified as *incorrect*. Their categorization of the SQL statements was based on Reisner’s categorization [7]. A few years later, Welty ran an experiment on a small group to test how assistance with error correction would affect user performance [9]. In that study, he categorized subject responses into “correct”, “minor error in problem comprehension”, “minor syntactic”, "complex errors", "group by", "s-type error", "incorrect" and "unattempted".

Buitendijk [10] introduced a classification of natural language questions, as well as possible errors within each class which resulted in four general groups of logical errors including "existence", "comparison", "extension" and "complexity". This categorization is not only based on SQL anomalies, but also focuses on user mistakes.

Smelcer [11] developed a model of query writing that integrated a GOMS-type analysis of SQL query construction with the characteristics of human cognition. This model introduced four common cognitive causes of JOIN clause omission and resulted in the categorization of common mistakes in writing SQL queries to "omitting the join clause", "AND/OR difficulties", "omitting quotes", "omitting the FROM clause", "omitting qualifications", "misspellings" and "synonyms". Brass [12] reports an extensive list of conditions that are strong indications of semantic mistakes. However, none of these studies analysed student mistakes in large datasets.

3. METHOD

3.1 Snapshot Collection

The data collected in this study forms a total number of ~161000 SQL SELECT statement snapshots from ~2300 students. Each snapshot is of one student attempt at a particular test question. The students in this study were all novice undergraduate students enrolled in an introductory database course. The tool used to collect the data is a purpose-built online assessment system named AsseSQL. Further details on the tool and how it was used to test the students can be found in prior publications [5, 6]. These snapshots were generated during supervised 50 minute online tests in which students attempted to answer seven SQL questions based on a given case study database. Students were provided with the case study, which included the description of the database, the Entity Relationship Diagram (ERD), and the CREATE statements corresponding to the database tables. Each question tests a student's ability to design a SELECT statement that covers a specific concept. Table 1 shows the concepts covered in the online test and statistics of snapshots related to each concept. A more detailed explanation on the nature of these concepts and their relative difficulty levels can be found in an earlier publication [13].

Table 1. Different SQL concepts and the number of snapshots.

Concept	Snapshot count
Group by with having	~32k (20%)
Self-join	~27k (17%)
Group by	~25k (15%)
Natural join	~24k (15%)
Simple subquery	~19k (12%)
Simple, one table	~18k (11%)
Correlated subquery	~16k (10%)

3.2 Snapshot Categorization

In order to produce the execution result of the collected snapshots, all snapshot SQL statements were re-executed in PostgreSQL and the output of each snapshot was obtained. Depending on the output returned by the PostgreSQL and the marking results of AsseSQL, each snapshot was tagged as *correct*, *syntactically wrong* or *semantically wrong*. We categorized each snapshot as a) correct if its result set was exactly the same as desired solution for the question corresponding to the snapshot, b) syntactically wrong when an error message was returned by the PostgreSQL, or c) semantically wrong when the execution of snapshot resulted in either an empty result set or a result set which was not exactly the same as the desired solution for the question corresponding to the snapshot. In this study, a student's snapshot is considered to be semantically incorrect if the output generated by the snapshot is different from the correct output. The categorization of the snapshots and its breakdown for each concept is shown in Table 2. While some of the snapshots in each level are correct, the majority of snapshots introduce an error (Figure 1). A detailed exploration of the reasons behind the syntactic errors is available in an earlier publication [14].

Table 2. Categorization of snapshots based on their output and their breakdown for different SQL concepts.

Concept	Correct	Syntactically wrong	Semantically wrong
Group by with having	4%	58%	37%
Self-join	2%	37%	61%
Group by	7%	63%	30%
Natural join	4%	64%	32%
Simple subquery	5%	61%	34%
Simple, one table	11%	48%	41%
Correlated subquery	6%	52%	42%
Among all snapshots	6%	54%	40%

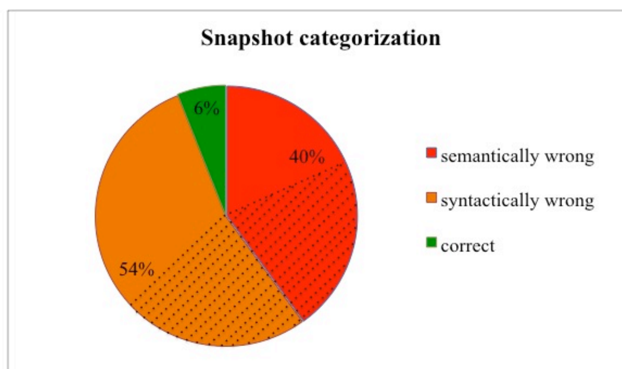

Figure 1. Categorization of collected SQL snapshots (N = ~161k). Snapshots in the hatched area were generated by students who subsequently fixed their error.

3.3 Semantic Mistake vs. Syntactic Error

As could be seen in Table 2 above, a considerable number of snapshots introduce an error, with the majority of snapshots falling into the syntactic error category. As the number of syntactic errors observed in students' attempts is higher than number of semantic mistakes, one could argue the importance of syntactic errors over semantic mistakes. However, in this study we chose to investigate the snapshots with semantic mistakes for the following reasons. Firstly, our analysis suggests that syntactic errors are more likely to be the result of lack of practice or carelessness. This is supported by the fact that among all syntactic error snapshots, ~69% of them are due to a typing errors in the SELECT statement (N = ~61000). These typing errors are due to wrong column names, wrong table names, or wrong syntax in one or more clauses of the SELECT statement. Excluding the syntax errors due to typing mistakes, the majority of mistakes made by novices are semantic. The second reason is related to the last snapshots generated by the unsuccessful students, which reflect the point where students stopped trying to get a query right. The number of students' last attempts with semantic mistakes is almost three times more than the number of last attempts that are syntactically incorrect (including typing mistakes). Another reason supporting our decision is that a SQL query that suffers from a syntactic error might already have a semantic mistake encapsulated in it. Our preliminary result suggests that a considerable number of *fixed* syntactic error snapshots are not correct as they include semantic mistakes (N = ~21000). Finally, the error code and the error message returned by PostgreSQL are usually enough to indicate the reason for a syntax error. In contrast, the output of a query that suffers from a semantic mistake is not as easily diagnosed as a syntactic error. This makes a semantic mistake much harder to fix.

3.4 Selection of Database Case Study

The set of ~161000 snapshots in this study are based on three different database case studies that are used in the online tests. The ERD structure as well as data complexity of these three databases are very similar. They consist of four to five tables, including one associate relation and three to four relationships, one of which is a unary relationship. The success rates of these case studies are only slightly different (Figure 2). The relative numbers of syntactic errors generated by students among these case studies are similar, however, the relative number of semantic mistakes made by students differs from one database case study to another. To make the result of our work less dependent on the comparative difficulty level of these case studies, we decided to limit our investigation dataset to the set of snapshots collected from only one database case study. We selected database case study 'Bicycle', which is based on a database from Post [15]. This database case study has the highest median success rate among different SQL query types, which reflects its lower level of difficulty relative to the other two case studies. For this case study we have ~45000 snapshots collected from ~700 students.

3.5 Primary Cause of Semantic Errors

According to Figure 1, the vast majority (94%) of snapshots do not generate the correct result, due to either a syntactic error or a semantic mistake. Almost half of these incorrect snapshots (46%) are from students who eventually were able to correct the errors and answer the question correctly. Figure 3 shows the frequency of incorrect snapshots **as a function of the attempt number** for these students who did eventually answer a question correctly. Among ~3200 cases where a student was able to answer a question correctly, 57% of students constructed at least one semantically incorrect SQL statement before they produced the correct SQL snapshot. Hence, we conclude that not all semantically wrong snapshots/student attempts are unfixable. This is also reported by Ogden *et al.* [16]. As a result, we elected to identify and investigate only those semantic mistakes that students were not able to correct. We limited our investigation to the set of snapshots (N = 551) that are final attempts with a semantic error. Those snapshots were generated by 321 students.

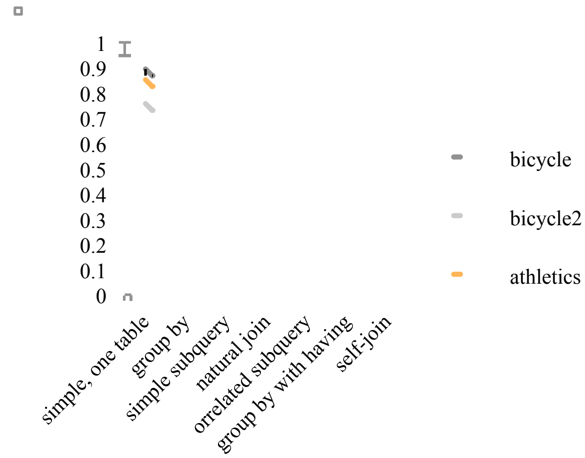

Figure 2. Success rates of different database case studies.

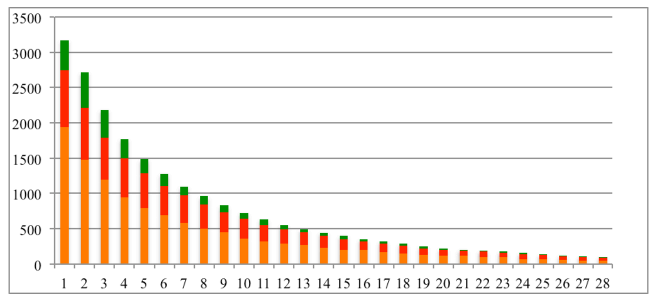

Figure 3. Distribution of correct (green), semantically incorrect (red) and syntactically incorrect (orange) snapshots, as a function of the number of attempts. X axis represents the n^{th} attempt and Y axis represents number of snapshots.

A snapshot may contain multiple semantic mistakes. We focused upon the primary semantic mistake. Our process for manually identifying the primary semantic mistake in the 551 snapshots is illustrated in the following example:

List the given name, family name, address, the employee ID and the number of managers of those employees who are less than 25 years and have been managed by more than one manager.

The correct query for such a question could be:

SELECT *EMP_ID, FNAME, LNAME, ADDRESS* , COUNT(*EMPID*) AS *NUMBER_OF_MANAGERS FROM EMP_DETAILS* NATURAL JOIN *EMP_MANAGER* WHERE *EMP_AGE* < 25 GROUP BY *EMPID* HAVING COUNT(*EMP_ID*) > 1;

A student's attempt for this question might be:

SELECT * FROM *EMP_DETAILS* WHERE *EMPAGE* < 24;

This SELECT statement has multiple semantic mistakes; however, the most important mistake is the absence of the EMP_MANAGER relation in this query. Even if all other aspects of the given query related to EMP_DETAILS table are corrected, this query will not produce the correct result set unless the EMP_MANAGER table is included in the query. Thus the principle semantic error is the omission of the table.

4. RESULTS

Queries of type *self-join* have the highest number of semantically wrong last attempts (178 snapshots). Questions of type *natural* *join*, *group by with having* and *correlated subqueries* have the second highest frequency. The *simple with one table* and *group by* categories have the lowest frequencies. In all categories, the main reason behind the semantic mistake is students' lack of skill in identifying the required type of query (as listed in Table 2). Table 3 reviews different clauses of a SELECT statement and the mistakes allocated to those clauses.

Table 3. Clause based categorization of principle semantic mistakes.

Clause	Mistake/s	Concepts
where (46%)	Missing/wrong condition	Simple, self-join, correlated subquery, join
from (26%)	Self-join not used	Self-join
having (13%)	Missing group by or having clause, use of wrong column	Having
order by (5%)	Missing order by clause, incorrect/incomplete column	Simple, group by
select (5%)	Missing/extra column	Simple, group by
group by (5%)	Missing group by clause, use of wrong column	Group by, group by with having

4.1 Simple Query with One Table
Construction of simple queries with one table is the first concept that novices learn at the authors' institute. The students investigated in this study have the highest success ratio for questions of this type. However, snapshots that belong to this category also suffer from semantic mistakes. The most common mistakes observed for this query type includes missing/unnecessary columns in a SELECT clause or an ORDER BY clause. The absence of an appropriate condition or the presence of irrelevant condition in a WHERE clause is also a common mistake. In a few cases, the GROUP BY clause was irrelevantly used. The majority of these snapshots fall into Reisner's category of "essentially correct", as they are mostly the result of carelessness.

4.2 Group By
Four common mistakes were identified, including mistakes related to the GROUP BY clause, SELECT clause, WHERE clause or ORDER BY clause. The most frequent mistakes include missing the GROUP BY clause entirely or including unnecessary columns in the GROUP BY clause. Wrong conditions in the WHERE clause and a missing ORDER BY clause were also among the most frequent mistakes for GROUP BY queries. Also, a missing aggregate function in the SELECT clause was a common mistake, which has direct correlation with the absence of a signal word in the question (e.g. "average" or "sum"). This was originally observed in Reisner's experiments and further supported by the findings of Ahadi et al. [13].

4.3 Simple Subquery
More than 80% of the investigated snapshots in this category did not use a subquery at all. This could indicate that either students were not able to identify the skill required to answer these questions, or they did not have the skill required to construct such a query.

4.4 Correlated Subquery
More than 70% of snapshots in this category lacked the structure of a correlated subquery. In other words, the identification by students of the need for a correlated subquery was the most common problem. A further 13% of snapshots followed the syntax of a simple subquery. The rest of the snapshots included a wrong condition in the WHERE clause of the inner query.

4.5 Join
More than 80% of the snapshots in this category lacked the JOIN clause, that is, either the ON clause or the JOIN clause in the WHERE clause was missing. This is perhaps explained by the fact that the use of "NATURAL JOIN" in a SQL SELECT does not require the specification of the joining keys in the query. As a result, students who wrote the join queries may have forgotten that, unlike the NATURAL JOIN, the INNER JOIN requires the indication of the keys involved in the join. Unnecessary use of aggregate functions and incorrect joining conditions in the WHERE clause were also observed.

4.6 Group By with Having
Around 75% of the snapshots in this category lacked a HAVING clause and 15% of them did not even include the GROUP BY clause. Upon closer inspection, we noticed that the condition that is supposed to appear in the HAVING clause was often mistakenly written in the WHERE clause. Around 15% of the snapshots included the wrong conditions in the HAVING clause.

4.7 Self-join
Around 75% of the snapshots in this category demonstrated that students did not understand that they needed a self-join. While the remaining snapshots had a self-join, the self-join lacked a condition in the WHERE clause by which the result set is limited to non-overlapping sections of the main table and its replicate. In rare cases, unnecessary/incorrect conditions were included in the WHERE clause.

4.8 Mapping Semantic Mistakes
We mapped the 551 snapshots to the categorizations provided by Reisner, by Welty and Stemple, by Welty, and by Buitendijk [7-10]. Note, however that most error categorizations in the literature are a mix of semantic and syntactic mistakes.

4.8.1 Psychological experiments and error categorization
Reisner has done most of the work in experimental analysis of learnability of SQL [7]. She categorized students' mistakes into "intrusion", "omission", "prior-knowledge", "data-type", "consistency" and "over-generalization". More than half of the observed semantic mistakes in the snapshots analyzed in this study are of type omission. This error usually happens when a signal word such as "average" is not given in the question. Our results support Reisner's experiment. While omissions were reported by Reisner, her observations were limited to GROUP BY omission, and also AND clause omissions. Welty's "incomplete" category is (to some extent) similar to Reisner's omission error category. Table 4 reviews different omission errors observed among the set of 551 snapshots in our study.

Table 4. Omission errors in semantically incorrect snapshots. The second number in each row represents the percentage among all omission errors.

category	Frequency
Omitting JOIN clause	218 (48%)
Omitting SUBQUERY	115 (25%)
Omitting HAVING clause	58 (12%)
Omitting ORDER BY clause	23 (5%)
Omitting GROUP BY clause	17 (3%)
Omitting aggregate function	8 (1%)
Omitting WHERE clause	6 (1%)
Omitting column in SELECT	3 (<1%)
Omitting DISTINCT	3 (<1%)
Omitting column in ORDER BY	2 (<1%)

4.8.2 Categorization based on ease of error correction

Welty's experiment classified subject's responses into multiple categories according to the amount of effort needed to correct their mistakes [9]. In our data, snapshots with a semantic mistake of type "minor spelling errors" were frequent. Most of the time, in our data, a misspelling resulted in a syntactic error – e.g. typing "SELCT" instead of "SELECT" – but there are cases where a misspelling results in a semantic mistake, for example, using the string 'hawaii' instead of 'Hawaii' when referring to data stored in a table.

A similar error to Welty's "minor spelling error" is the "overgeneralization" error proposed by Reisner. For a given question "List the names of employees…", an overgeneralized solution would be:

SELECT *NAMES* FROM *EMPLOYEE … etc.*

The overgeneralization error happens when the information given in the question is directly extracted and used to construct the query, in the case above the *names* and *employees*. Reisner's overgeneralization category overlaps with Buitendijk's "comparisons" error category.

In some cases, snapshots of this type appeared immediately after a syntactically wrong attempt where the name of the columns or the tables were directly extracted from the question itself. Reisner classifies this kind of syntactic error as "intrusion error". The result of syntactic error analysis investigated by Ahadi *et al.*[14] reported that this type of mistake was common. In that study, syntactic errors due to prior knowledge – e.g. using AVERAGE instead of AVG – were also reported to be common.

Some semantic mistakes fall into Welty and Stemple's "major substance" error [8]. This error happens when the query is syntactically correct but the query answers a different question. Such errors are hard for students to detect and fix. Examples of these types of errors are a missing second column in an ORDER BY clause, an unnecessary condition in a WHERE clause, or the use of an incorrect aggregate function. This error was not common in our snapshots.

4.8.3 Natural Language based Categorization

Proposed by Buitendijk [10], this categorization of logical errors introduces "existence", "comparison", "extension" and "complexity" errors. Among our data, over one quarter of the semantic mistakes fall into the complexity category. An example of such cases is when more than one subquery needs to be included in the query.

Semantic mistakes due to the complexity of the written query are hard to detect. A query is regarded as complex if the translation the natural language question to the SQL answer is difficult. Interestingly, there is no relationship between the complexity of the given question and its corresponding answer. For example, longer questions do not necessarily require a complex query. However, longer queries are usually more complex and as a result more likely to contain errors.

5. DISCUSSION

Ogden *et al.* [16] categorized the knowledge of query writing into knowledge of the data, knowledge of the database structure and knowledge of the query language. The lack of knowledge of the first two categories is best reflected in syntactic errors, particularly those errors due to misspelling or referring to undefined columns or tables. On the other side, lack of knowledge of the query language is better reflected in semantic mistakes. Our findings suggest that the primary reason behind semantic mistakes is students' poor skill in selecting the right technique to design the query needed to answer the test question. The presence of a signal word such as "average" in the question seems to be helpful; however, students' dependency on a such "clue" in the question is not ideal.

Semantic mistakes also have implications for students' development in writing queries. Reisner's [7] model for the development of a query consists of three phases; generation of a template, transformation of English words to database terms, and insertion of database terms into the template. However, we noticed that the majority of our students who abandoned the question due to a semantic mistake had problems with the generation of the initial template. The first step in generating the template is to identify which technique – e.g. natural join, simple subquery, self-join, etc. – is most suitable. Our results suggest that there should be a greater emphasis on this matter when teaching students how to write SQL queries.

Our results show that the majority of semantic mistakes occur in the WHERE clause of the SELECT statement. This error may occur when the capacity of a student's working memory is surpassed [17]. For example, a high number of conditions in the WHERE clause could result in working memory overload, especially when many WHERE conditions are required to join two or more tables. This has been previously shown in the experiment performed by Smelcer [11]. This problem could perhaps be avoided by putting greater emphasis in teaching on following a systematic, step-by-step procedure in segmenting the question and formulating the correct answer in SQL.

6. CONCLUSION

This study attempts to both qualitatively and quantitatively investigate the semantic mistakes made by students. The results of our analysis show that syntactic mistakes are more common than semantic mistakes. However, semantic mistakes are much harder to correct, even among successful students. Our findings show that the majority of semantic mistakes are of type omission, indicating that students have difficulty with selecting the correct type of query, and they also lack a systematic approach to formulating the query.

The majority of omission errors happen in queries that require a JOIN, or a subquery, or a GROUP BY. This has implications for the way that we teach SQL query design, in that we should emphasize techniques that deal with the identification of the type of query necessary to return the requested information, e.g. a JOIN or a subquery. Furthermore, we believe that the selection of the right terminology by the teacher in articulating the question will decrease the chance of students' semantic mistakes that are due to complexity or the comprehension of the question. We are carrying out additional research to confirm this assertion.

7. REFERENCES

[1] Brusilovsky, P., Sosnovsky, S., Lee, D., Yudelson, M., Zadorozhny, V., and Zhou, X. (2008). *An open integrated exploratorium for database courses*. ITiCSE '08. pp. 22-26. http://doi.acm.org/10.1145/1384271.1384280.

[2] Brusilovsky, P., Sosnovsky, S., Yudelson, M. V., Lee, D. H., Zadorozhny, V., and Zhou, X. (2010) *Learning SQL programming with interactive tools: From integration to personalization*. Trans. Comput. Educ. 9, 4, Article 19 (January 2010). http://doi.acm.org/10.1145.1656255.1656257

[3] Mitrovic, A. (1998) *Learning SQL with a computerized tutor*. SIGCSE '98, pp. 307-311. http://doi.acm.org/10.1145/274790.274318

[4] Mitrovic, A. (2003) *An intelligent SQL tutor on the web.* Int. J. Artif. Intell. Ed. 13, 2-4 (April 2003), pp. 173-197.

[5] Prior, J., and Lister, R. (2004) *The Backwash Effect on SQL Skills Grading*. ITiCSE 2004, Leeds, UK. Pp. 32-36. http://doi.acm.org/10.1145/1007996.1008008

[6] Prior, J. (2014) *AsseSQL: an online, browser-based SQL Skills assessment tool*. ITiCSE 2014. Pp. 327-327. http://doi.acm.org/10.1145/2591708/2602682

[7] Reisner, P. (1977) *Use of psychological experimentation as an aid to development of a query language*. IEEE Trans. Softw. Eng. SE-3, 3, 218-229. http://doi.acm.org/10.1145/1103669.1103673

[8] Welty, C., and Stemple, D. W. (1981) *Human factors comparison of a procedural and a nonprocedural query language.* ACM Transactions on Database Systems (TODS) 6.4:626-649. http://doi.acm.org/10.1145/319628.319656

[9] Welty, C. (1985) *Correcting user errors in SQL*. International Journal of Man-Machine Studies 22(4): 463-477.

[10] Buitendijk, R. B. (1988) *Logical errors in database SQL retrieval queries*. computer Science in economics and management 1, 79-96. http://doi.acm.org/10.1007/BF00427157

[11] Smelcer, J. B. (1995) *User error in database query composition*. Int. J. Human-Computer Studies 42, 353-381. http://doi.acm.org/10.1006/ijhc.1995.1017

[12] Brass, S. and Goldberg, C. (2006) *Semantic error in SQL queries: A quite complete list*. The Journal of Systems and Software 79, 630–644. http://doi.acm.org/10.1016/j.jss.2005.06.028

[13] Ahadi, A., Prior, J., Behbood, V., and Lister, R. (2015) *A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries*. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, pp. 201-206. ACM, 2015.

[14] Ahadi, A., Behbood, V., Vihavainen, A., Prior, J., & Lister, R. (2016, February). *Students' Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students' Success.* In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 401-406). ACM.

[15] Post, G. V. (2001) *Database management systems: designing and building business applications*. McGraw-Hill.

[16] Ogden, W. D., Korenstein, R., Smelcer, J. B. (1986) *An Intelligent Front-End for SQL*, IBM, San Jose, CA.

[17] Miller, G. A. (1994). *The magical number seven, plus or minus two: Some limits on our capacity for processing information.* Psychological review, *101*(2), 343.