ISSN: 2560-5429 Journal of Applied Technical and Educational Sciences http://jates.org/
Vol. 15, No. 1: 406 https://doi.org/10.24368/jates406 1-17

Impact of Educational Escape Games on Students' Semantic Lexicon in Environmental Science: A Word Association and AI-Based Analysis

Mihály Kovács [1*]

[1*] *Observatory and Science Experience Center, Eszterházy Károly Catholic University, Eszterházy sq. 1., Eger 3300, Hungary, kovacs2.mihaly@uni-eszterhazy.hu*

Abstract: *Gamification, including escape games, has proven to be an effective tool for changing environmental behaviour. However, little research has been done on the impact of this game type on students' conceptual networks, particularly in relation to environmental science and chemistry. Therefore, this study examines the semantic lexicon of primary and secondary school students using word association tests with selected call words from these subjects. Students were asked to provide answers immediately before the game and exactly one week later. The pre- and post-tests were analysed using the rather robust but complicated Garskof-Houston formula and the results were compared to the analysis of a smaller generative AI running on a laptop to find a more user-friendly way to evaluate this type of survey. The Garskof-Houston formula showed that the relations of the words "mixture", "precipitation" and "rain" changed the most, but the smaller, and therefore presumably weaker, LLM found the most changes for the relations of the words "compound", "precipitation" and "rolling". Since the meaning of a concept can be defined as a list of words with which it is associated, these results suggest that learners' concepts have changed, which is quite important from the perspective of constructivist learning theory. However, further research is needed on the use of LLMs for this type of evaluation.*

Keywords: *Gamification; Escape Game; Conceptual Change; AI Analysis; Chemistry*

1. Introduction and literature review

Gamification "is a careful and considered application of game thinking to solving problems and encouraging learning using all the elements of games that are appropriate." (Kapp, 2012) This definition also includes serious games like educational escape games. These are "live-action team-based games where players discover clues, solve puzzles, and accomplish tasks in one or more rooms in order to accomplish a specific goal (usually escaping from the room) in a limited amount of time." (Nicholson, 2015)

Different game types were tried in the field of environmental education (Hallinger et al., 2020) and e.g. gamification in general (Charkova, 2024) and also escape games (Chang, 2019) proved to be effective tools to form participants' environmental behaviour. However, this study examined the effects of this game type to the players' knowledge, namely their semantical lexicon.

Semantical lexicon is part of the mental lexicon, which is a storage system in the long-term memory. It contains the elements (word or concepts) and rules of the language which are related to each other. This can be modelled as a network which nodes are the concepts (Carey, 2000; Gósy, 2005). The mental lexicon may contain different strengths and types of relatedness, e.g. rhymes. The semantical lexicon is its subnetwork narrowed down to the connections based on the interpretation of the elements. Therefore, Shavelson defined the meaning of the concepts as the list of those elements, with which it is connected. (Shavelson, 1972)

Piaget (Piaget, 1964) named the process which is taking place in the learner's mind construction, because they build actively they own knowledge, which is the basic statement of constructivist learning theory. Due to individual processing, every human being has its own cognitive structure, therefore a unique meaning of the same concept (Nahalka, 2002), and Shavelson's definition makes it possible to examine this personal construction. One possible way to identify a part of the semantical lexicon is the word association test evaluated with Garskof and Houston's relatedness coefficient (RC) (Garskof & Houston, 1963; Shavelson, 1972; Tóth, 2024) and if a concept changes, its connections will change, too. Moreover, the average of RCs can describe a class's cognitive structure (Tóth, 2024) and based on this examination of conceptual change with statistical analysis is possible.

This is crucial, since conceptual change is the key moment of learning in constructivist theory. This is a conflict caused by new experiences. During this process the students question their current theories and develop new, more adaptive concepts, which they are going to use at least in certain circumstances. (Nahalka, 2002) However, it is fairly common in environmental education, that students have no naive theories and the real question is if they can use their knowledge in a new situation. (Robertson, 1994; Robottom, 2004)

In preparing this study, I used a widely applicable method that helps to recognize and visualize conceptual changes during short educational interventions. This method uses word association tests to analyse changes in students' semantic lexicon, i.e., how individual concepts are related to each other in their long-term memory. This provides a deeper understanding of the learning process, especially in the context of constructivist learning theory.

2. Research aims and questions

The research question of this paper is whether it is possible to achieve changes in the students' semantic structures with the help of escape games, if the examined concepts are selected from the field of chemistry and environmental sciences. For the purposes of the study, the following

words were selected for 7th grade: mixture, compound, and wastewater, while for 9th grade: precipitation, rain, fluid, rolling, and water cycle. Based on this, the research question can be formulated more specifically:

RQ1: How much does the 7th graders' semantic network change in relation to the concepts of "mixture," "compound," and "wastewater" as a result of the escape game?

RQ2: How much does the 9th graders' semantic network change in relation to the concepts of "precipitation," "rain," "fluid," "rolling," and "water cycle" as a result of the escape game?

Since, from a theoretical point of view, educational escape games can be linked to constructivist learning theories (Nicholson, 2018; Zhang et al., 2018), the meaning of concepts related to the game is likely to change during this activity. Hence, the research hypothesis is the following: as a result of playing escape games, changes can be observed in the strength of the connections among the selected concepts in the players' semantic lexicon.

This hypothesis was tested using statistical methods by calculating the t-test for each pair of call words. Thus, the hypotheses related to questions RQ1 and RQ2 can be reformulated as follows:

H1: For at least one pair of the call words "mixture", "compound", and "wastewater" the Garskof-Houston coefficient changes significantly between the pre-test and post-test.

H2: For at least one pair of the call words "precipitation", "rain", "fluid", "rolling", and "water cycle" the Garskof-Houston coefficient changes significantly between the pre-test and post-test.

However, calculating the Garskof-Houston coefficient is a time-consuming process, so the question came up whether word association tests could be analysed more simply but still reliably to make them easier to use in teachers' everyday work. Since the information is text-based, the use of a large language model (LLM) in the analysis seemed an interesting possibility. Since I did not ask for permission to provide the students' answers to an AI that might learn from them, for ethical reasons I could only use a weaker model run on my own computer, which is not expected to give very accurate answers. Therefore, hypothesis testing was not performed; only the theoretical possibility of this method was analysed by comparing the LLM's analyses with the calculations. This can be formulated as a research question in the following form:

RQ3: How reliable is the generative AI (LLM) analysis compared to Garskof–Houston RC?

3. Methodology

A quasi-experimental research design was used with pre- and post-testing to decide whether the students' semantic lexicon changed significantly during an escape game. According to theory, this type of game can be linked to constructivist learning theory (Nicholson, 2018; Zhang et al., 2018), meaning that when used for educational purposes, changes should occur in the students' conceptual system, which fits well to their semantic lexicon.

The students completed the pre-test immediately before the game, the post-test was written one week after the game, as long-term memory changes were relevant regarding the research question. No homework was given to the students for this week in order to measure only the impact of the game. Hopefully, this one-week break also minimized the so-called priming effect, which means that the answers given in the pre-tests should not have a major impact on the responses given in the post-tests, as the students were studying every other subjects during this time.

3.1. Participants

The examinations took place in 7^{th} and 9^{th} grades, with participants selected by convenience sampling from rural and urban primary schools and urban high schools. Players who completed only either the pre-test or the post-test were omitted from the statistical analysis because both results are needed to calculate the paired sample t-test. In the end, the sample included N_7=68 students in grade 7 and N_9=98 students in grade 9.

This study received Eszterházy Károly Catholic University ethics approval (reference number: RK/144/2025). All methods were carried out in accordance with relevant guidelines and regulations

3.2. Educational games

The 7th graders played an escape card game about categorizing matter and its connection with wastewater. The puzzles were in sequential order, see Fig. 1. Every card had a title created from the number of the exercise and a key word from the story. There were cards with elements of the story, with lecture materials and with exercises. The players could draw the cards of the next puzzle from their own pile only if they solved the current exercise correctly. They could check their results on answer cards set on the teacher's desk.

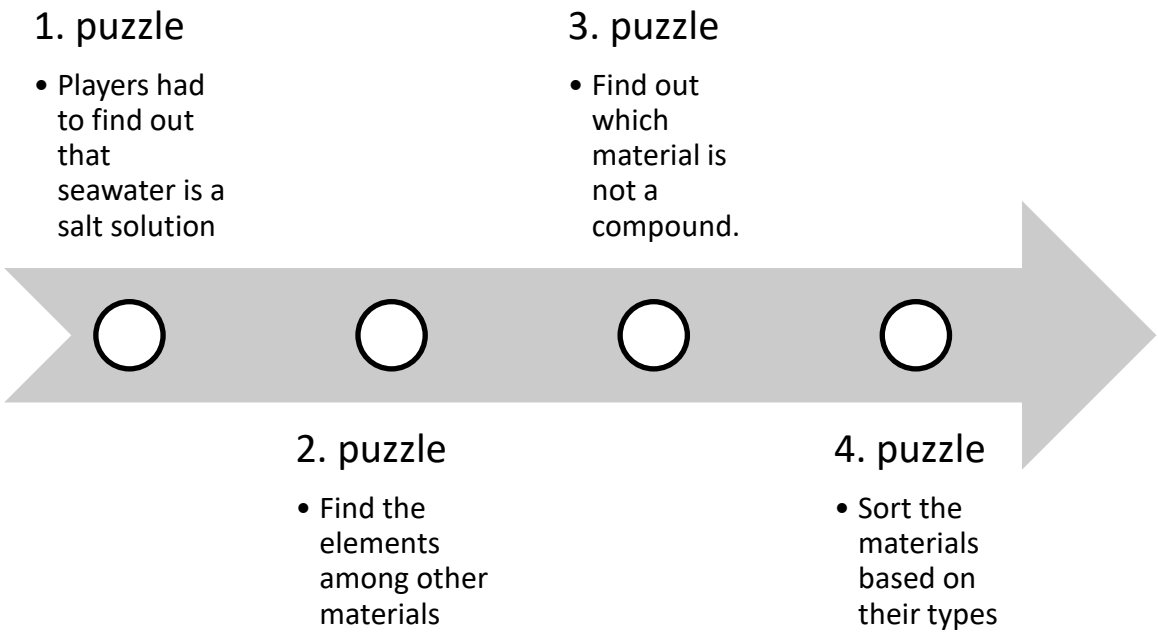

Fig. 1. Structure of the 7th graders' game

The 9th graders played a game like an escape book. Every riddle was on a separate page with some lecture materials and with the next part of the story. The exercises were also in sequential order, see Fig. 2, in this case, there were also answer cards on the teacher's desk to check their solutions. If their answer was correct, they could read the next page. The topic of this game was the water cycle and the state of matter.

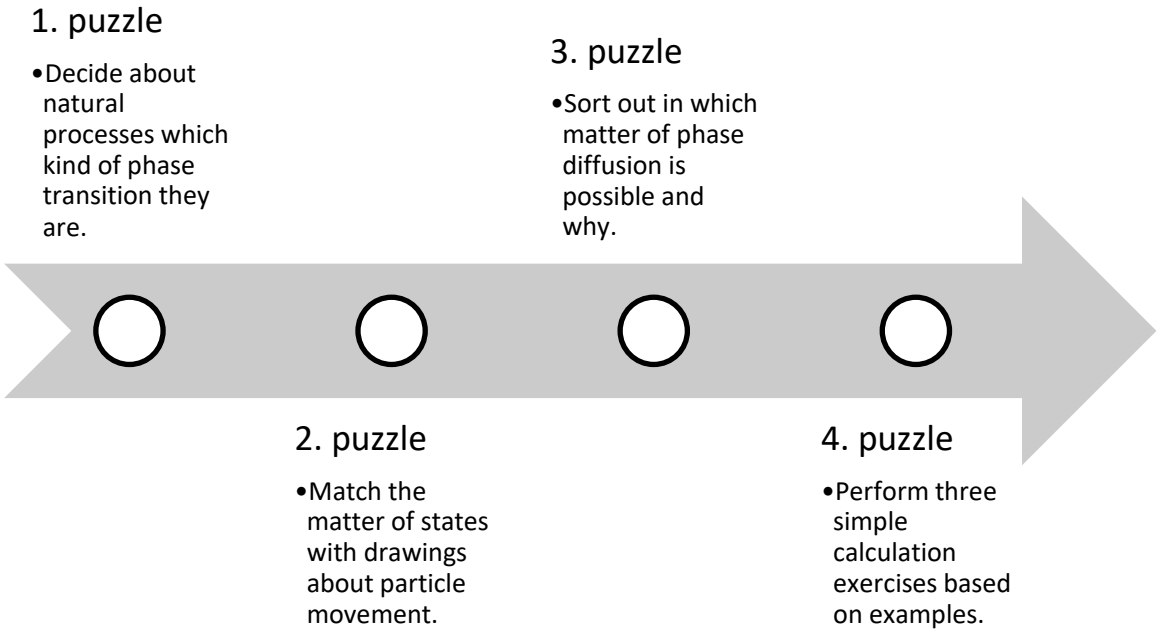

Fig. 2 Structure of the 9th graders' game

These sorting and selecting type of exercises hopefully triggered interaction, sometimes debate, among the players which could lead to cognitive conflict. (Bélanger, 2011) There were also short debriefs after both games, so the students had the opportunity to ask the teacher, who could also emphasize some points from the material, if they wanted to. However, they could not help during the game except for technical issues.

3.3. Word association tests

Word association tests contained call words, and the participants had to provide those words that come to their mind. A list of a call word and its corresponding associations in the order given by the respondent is called an associative meaning (i.e.: pink, magenta, colour, flower, flamingo is the associative meaning of pink).

During the test, they could work on only one word at a time, for 1 minute in this study, and after that, they had to proceed to the next buzzword and could no longer return to the previous ones. It was completed either on Microsoft Forms or on paper in this research, and the teacher measured the time. The data were anonymized right after pairing the responses to the pre- and post-test. The names were replaced with random numbers during anonymization, then the spreadsheet was sorted in ascending order based on these numbers.

Before the test, the teacher emphasized that the students should answer like an ecology specialist and they should not use complete sentences, only words or phrases. However, many students gave sentences at the end, so the answer needed to be coded, which was executed like in Daru and Tóth (2014). They cut the sentences into words and phrases but deleted the meaningless and irrelevant words like articles.

After that, negative words were also deleted, because they suggested the direction of the association and not the associated word. All the typos were corrected, too. When a student listed a set of adjectives in relation to a noun, those phrases were exceptionally split into parts. Since Hungarian is an agglutinating language and students used different kind of suffixes without meaning another word, in most cases, they were deleted, except when their changed the grammatical category of the word. Finally, the respondents gave the same associations twice in some cases, thus the second one was deleted.

The 7th graders' game was about wastewater and categorizing matter, so the selected concepts were wastewater, compounds and mixtures. The 9th graders' game was about phase change and

water cycle, so the selected concepts were water cycle, precipitation, rolling, fluid, rain. Rolling referred to the movement of particles in fluids.

3.4. Garskof-Houston relatedness coefficient

The Garskof-Houston formula (Garskof & Houston, 1963; Tóth, 2024) assigns a number called relatedness coefficient (RC) from the interval 0 to 1 to each associative meaning pair. The RC is 1 only in that case if the two buzzwords are synonyms, which means the first association is the other buzzword in both cases, and all other associations are the same in the exact same order. For example, $RC_{magenta,\ pink}$=1, if the associative meaning of magenta is pink, colour, flower, flamingo and the associative meaning of pink is magenta, colour, flower, flamingo. The result is 0 if there are no common words in the associative meaning of the two buzzwords. Since most papers contain some ambiguity or simplification it is useful to explain the equation in detail.

To calculate the RC, we needed to assign a rank number to all words in associative meanings of the two examined call words. The rank number of the buzzwords is equal to the number of elements of the associative report with the higher number of items. Then we had to go through the words of each associative meanings one by one, the current word was always assigned a rank number one lower than the previous one, as can be seen in Table 1.

Table 1. How to assign rank numbers to associative meanings

1st call word	rank number	2nd call word	rank number
pink	5	*blue*	5
magenta	4	colour	4
colour	3	sky	3
flower	2	pink	2
flamingo	1		

After this, we can understand these formulas which are:

$$RC = \frac{\bar{\mathbf{A}} \cdot \bar{\mathbf{B}}}{n^p(n-1)^p + (n-1)^p n^p + (n-2)^p(n-2)^p + \cdots + 1} \tag{1}$$

$$RC = \frac{\bar{\mathbf{A}} \cdot \bar{\mathbf{B}}}{\mathbf{A} \cdot \mathbf{B} - [n^p - (n-1)^p]^2} \tag{2}$$

where:

- n is the biggest rank number
- p is a power weight which had to be a natural number. In education, p=1 is the most frequent choice (Cardellini, 2008), however if the number of associations are maximized and/or the participants had to ranking their associations, p>1 can be a better

choice. This can be an interesting option, especially in the case of online data collection as maximum word number is easier to implement on most online form software than time limits for every question.

- $\overline{\mathbf{A}}$ vector contains the rank number of those words in the first associative meanings which are also listed in the second associative meaning. In the meantime, $\overline{\mathbf{B}}$ vector contains the rank number in the second associative meaning of the same words in the same order than $\overline{\mathbf{A}}$.
- **A** and **B** vectors are identical: $[n^p \quad n-1^p \quad \cdots \quad 1^p]$. It is easy to prove with algebraic identities that the two denominators are the same.

Equation (2) is most frequently used in literature however version (1) makes it easier to understand thus it is used in this paper. Its denominator contains a scalar product of an ideal situation. In this case, the respondent uses the two words as total synonyms like in previously mentioned examples about pink and magenta. Hence, the buzzwords rank number is n in one case, when we create the $\overline{\mathbf{A}}$ vector and *n*-1 in the other case ($\overline{\mathbf{B}}$ vector), and the ranks of all other words are the exact same. This makes RC smaller than 1, with other words normalizing it. For example, based on Table 1 with the choice of p=2, we got the equation (3):

$$RC = \frac{[5^2 \quad 3^2]\cdot\begin{bmatrix}2^2\\4^2\end{bmatrix}}{5^2\cdot4^2+4^2\cdot5^2+3^2\cdot3^2\ldots+1\cdot1} = \frac{5^2\cdot3^2+2^2\cdot4^2}{2\cdot5^2\cdot4^2+3^2\cdot3^2\ldots+1\cdot1} \approx 0.32 \tag{3}$$

Calculating RC for all student's all buzzword pair is quiet time consuming; therefore, this study used a previously validated LibreOffice macro (Kovács, 2024) for this purpose.

3.5. Statistical methods

Since the possible values of RC come from the interval 0-1, it can be analysed with a t-test if the data are normally distributed. The normality of the variables was checked with the Anderson-Darling test. The alpha level was chosen to be 0.05 for all statistical analyses. BlueSky Statistics 10.3.2 software was used to perform statistical analysis in this study. (Muenchen, 2023) The calculations were also checked using Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) The results of these calculations are presented in the conclusion, along with the limitations.

3.6. Generative AI Analysis

There may be other ways to make the evaluation of word association test simpler, for example, generative AIs. These can be used for different purposes in education, not only to help the

learners (e.g. personalized learning) but also to make the teachers' work easier since they can reduce their workload. (Iyamuremye et al., 2024; Mohebi, 2024) Thus, an LLM (Llama 3 8B in GPT4All software) was used in this research for this purpose, and its results were compared to the Garskof-Houston method. The advantage of this software is that it is running on the researcher's computer, so it is safer in the field of data protection than its alternatives, however it is much slower and weaker than the more well-known ones. Since the goal was to make teachers' work easier, I used LLM with default settings in this research, which means that temperature was rather high, equal to 0.7.

The following prompt was used on the raw data, without any coding or correcting the typos: "Answer my questions as an education specialist. You should evaluate a word association test completed by a group of students. The first call word was mixture. The following words came to the mind of the class members. The answers of each student are separated by semicolons: … What concepts would you connect the word "mixtures" with if you should create a mind map of the students' answers?" After this, the same question was repeated for all buzzwords, separately for the pre- and the post-test. Since this LLM was a smaller one run on my computer and the temperature was set quite high, the level of its answer was varied. To make the answers comparable, I regenerated them if they did not contain categories and some words from the students' associations as an example of the categories.

This process led to a very complex network, where the nodes are the buzzwords, the categories and the examples. There were different types of connections, i.e. the strongest is when the category is also a call word, and the weakest is when the example is shared. Moreover, there were parallel connections, too. They have been simplified for greater clarity, bold lines in the graphs indicate that there was minimum a common category between the two call words and thin lines mean any other type of detectable relationship.

4. Results

Before the calculation of the RCs, I coded the answers of the tests twice with two weeks difference. After comparing the two results and making some corrections in case of 9 graders, 81.25% of the codes were identical, while in case of 7 graders, 88,7%. I accepted them, and used the improved first version, because coding of sentences was more consequent. The means of the RC values in pre- and post-tests are in Table 2 and 3.

Table 2. Means and standard deviations (SD) of 7th graders' relatedness coefficient values

	mean ± SD pre-test	mean ± SD post-test
mixture-compound	0.1±0.17	0.22±0.23
mixture-wastewater	0.01±0.03	0.06±0.17
compound-wastewater	0.01±0.04	0.05±0.05

Table 3. Means and standard deviations (SD) of 9th graders' relatedness coefficient values

	mean ± SD pre-test	mean ± SD post-test
precipitation-rolling	0±0.03	0±0.02
precipitation-fluid	0.11±0.13	0.15±0.14
precipitation-rain	0.09±0.14	0.1±0.15
precipitation-water cycle	0.07±0.13	0.1±0.16
rolling-fluid	0±0.02	0±0.02
rolling-rain	0±0	0.01±0.05
rolling-water cycle	0±0	0±0.03
fluid-rain	0.15±0.17	0.22±0.21
fluid-water cycle	0.06±0.09	0.07±0.12
rain-water cycle	0.13±0.17	0.19±0.2

Before the calculation of the t-test, normality of RC values was also checked for every buzzword pair with the Anderson-Darling test. Based on these calculations, they can be handled as normal variables, the exact results are in Table 4 and 5.

Table 4. Results of normality tests for 7th graders relatedness coefficients
($^{*}p<0.05$ $^{**}p<0.01$ $^{***}p<0.001$)

	A – pre-test	A – post-test
mixture-compound	11.4466***	3.7987***
mixture-waste water	24.6989***	19.5812***
compound-waste water	24.2646***	23.0988***

Table 5. Results of normality tests for 9th graders relatedness coefficients
($^{*}p<0.05$ $^{**}p<0.01$ $^{***}p<0.001$)

	A – pre-test	A – post-test
precipitation-rolling	34.9539***	35.1575***
precipitation-fluid	8.7939***	5.2396***
precipitation-rain	12.4284***	9.2954***
precipitation-water cycle	14.0088***	12.6949***
rolling-fluid	36.6896***	36.0987***
rolling-rain	36.6896***	33.7954***
rolling-water cycle	4.0841***	33.8596***
fluid-rain	6.1437***	3.9610***
fluid-water cycle	15.5334***	15.6805***
rain-water cycle	9.1762***	4.0841***

Table 6 presents the results of the t-tests, which showed significant changes in the case of mixture-compound (t=-4.2403, p<0.001, d=-0.5142) and mixture-wastewater (t= -2.3395, p=0.0223, d=-0.2837) call word pairs in case of 7th graders. While in case of 9th graders, t-tests showed significant changes for the precipitation-fluid (t=-2.0586, p=0.042, d=-0.2101), fluid-

rain (t=-3.3560, p=0.001, d=-0.3425) and rain-water cycle (t=-2.6636, p=0.009, d=-0,2719) buzzword pairs, as you can see in Table 7.

Table 6. Results of paired sample t-tests for 7th graders relatedness coefficients (*p<0.05 **p<0.01 ***p<0.001)

			confidence interval (95%)	
	t	Cohen's d	low	high
mixture-compound	-4.2403***	-0.5142	-0.7713	-0.2614
mixture-waste water	-2.3395*	-0.2837	-0.5291	-0.0405
compound-waste water	-0.3916	-0.0475	-0.2873	0.1919

Table 7. Results of paired sample t-tests for 9th graders relatedness coefficients (*p<0.05 **p<0.01 ***p<0.001)

			confidence interval (95%)	
	t	Cohen's d	low	high
precipitation-rolling	0.2387	0.0244	-0.1767	0.2255
precipitation-fluid	-2.0586*	-0.2101	-0.4140	-0.0073
precipitation-rain	-1.5430	-0.1575	-0.3602	0.0444
precipitation-water cycle	-1.6357	-0.1669	-0.3699	0.0351
rolling-fluid	-0.5985	-0.0611	-0.2625	0.1400
rolling-rain	-1.7806	-0.1817	-0.3850	0.0205
rolling-water cycle	-1.7867	-0.1824	-0.3856	0.0199
fluid-rain	-3.3560**	-0.3425	-0.5504	-0.1365
fluid-water cycle	-0.9645	-0.0984	-0.3003	0.1029
rain-water cycle	-2.6636**	-0.2719	-0.4774	-0.0678

The following section presents the mind maps generated with the help of AI. Graphs of Fig. 2 constructed based on the LLM's analysis of 7th graders' answers show that the relationship between the mixture-compound and mixture-waste water call word pairs changed from the pre-test to the post-test.

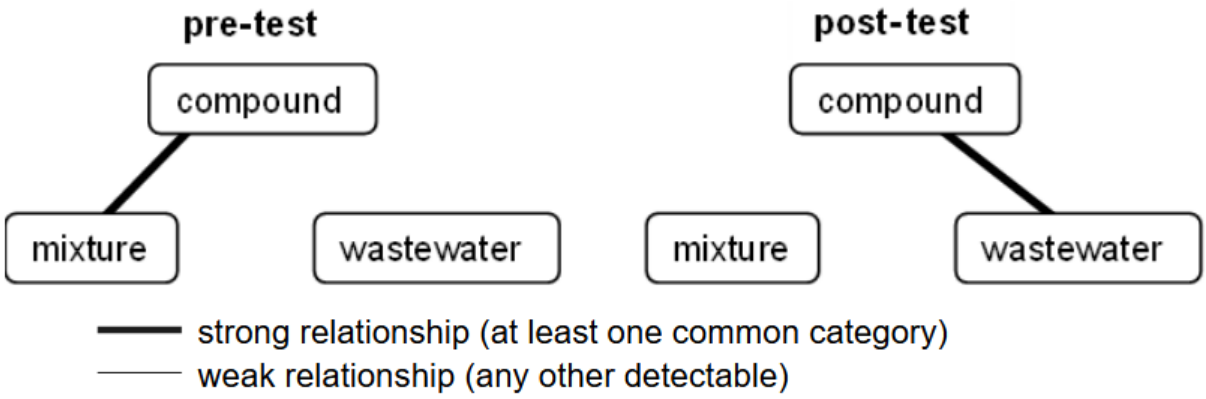

Fig. 2. Graphs constructed based on the LLM's analysis of 7th graders associations.

The 9th graders graph presented on Fig. 3 shows that precipitation-rolling connections disappeared, but rolling-rain and rolling-fluid turned up. Connections with precipitation also became weaker.

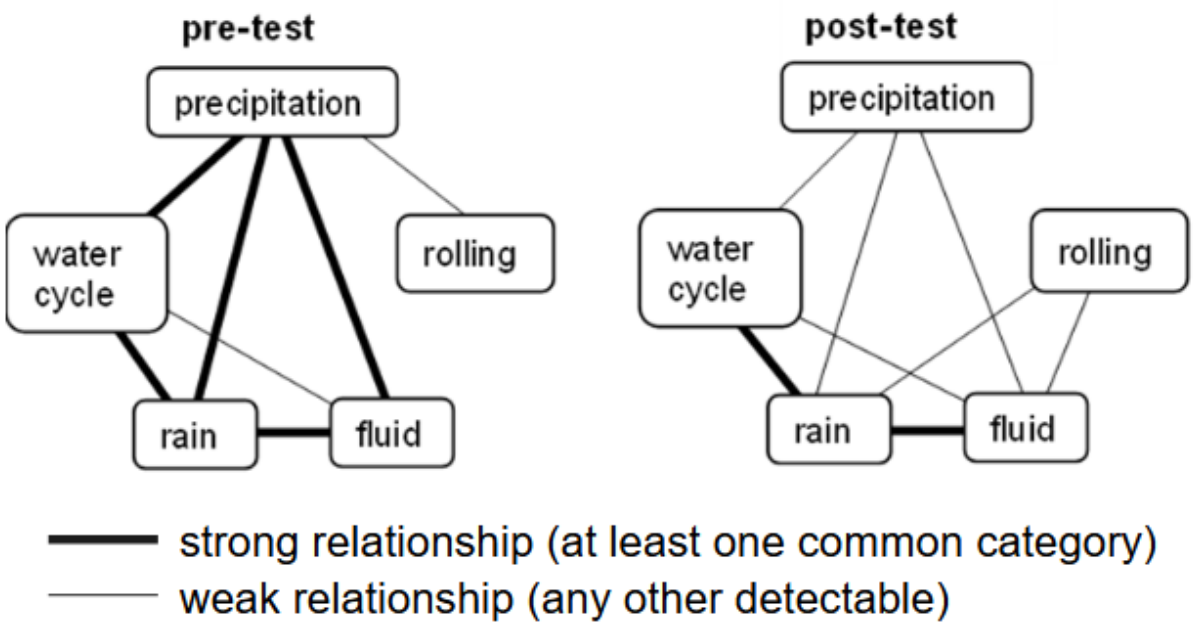

Fig. 3. Graphs constructed based on the LLM's analysis of 9th graders associations.

5. Discussion

The statistical analysis indicates that both the 7th graders' and the 9th graders' semantical lexicon has changed significantly, which means based on Shavelson's definition (Shavelson, 1972) that their understanding of the examined concepts changes. For 7th graders, the meaning of the word "mixture" changed the most, since both RC values changed significantly (mixture-compound: $t=-4.2403$, $p<0.001$, $d=-0.5142$; mixture-wastewater: $t= -2.3395$, $p=0.0223$, $d=-0.2837$). Based on this, we can accept H1, meaning that there was a significant change in the students' semantic lexicon in the case of "mixture," "compound," and "wastewater".

For 9th graders the buzzwords "rain" has changed the most, two RC values changed significantly (fluid-rain: $t=-3.3560$, $p=0.0011$, $d=-0.3425$; rain-water cycle: $t=-2.6636$, $p=0.0091$, $d=-0.2719$). Besides that, the RC changed significantly in the case of another pair of call words, namely precipitation-fluid ($t=-2.0586$, $p=0.0423$, $d=-0.2101$). This means that we can accept H2, too, meaning that there was a significant change in the students' semantic lexicon in the case of "precipitation," "rain," "fluid," "rolling," and "water cycle".

In case of 7th graders, the effect size ($d=-0.5142$) of the mixture-compound call word pair was greater than 0.39, which was the average obtained by Tamim et al. (2011) in their meta-analysis in the field of education, but the mixture-waste water pair ($d=-0.2837$) fell below average. In the case of the 9th graders, the effect size of fluid-rain pair ($d=-0.3425$) was close to average, but the other pairs fell below average ($d=-0.2719$ and $d=-0.2101$).

Turning to RQ3, LLM's analysis also points in this direction, but with some serious errors. It found the connection between the mixture and the compound in the case of 7th graders, and it realizes its change but missed the direction. It found that the meanings of precipitation and rain changed a lot, but missed the direction in this case, too. Besides, it highly overestimated the change in the meaning of rolling.

However, due to ethical reasons, a small model was used in this study, therefore its results cannot be so accurate as it would be with the most popular ones. Hence, more research is needed about the question of using LLM to analyse word association tests, and to examine someone's semantical lexicon, even on the basis of other data types (e.g. texts).

Beyond empirical results, the findings described above were discussed in a framework that can be applied in classroom work or in another research. The use of word association tests as pre- and post-tests, followed by their evaluation, can provide educators with a practical tool for understanding changes in students' conceptual networks. We discussed two methods of evaluation: statistical analysis of the Garskof–Houston coefficients obtained in this context, and evaluation using LLM. This allows for a more complex understanding of the changes taking place in students during the learning process.

6. Conclusion

Based on these considerations, educational escape games can significantly transform how students understand scientific concepts. The results, especially the more robust statistical analysis of Garskof-Houston RCs, suggest that some of the students either started to construct new concepts or at least to refine some existed ones, which is a main goal in Robertson's and in Robottom's theory (Robertson, 1994; Robottom, 2004), or some of them started the process of so called conceptual change, which is a main goal of constructivism in general (Nahalka, 2002). This means, that educational escape games can be effective tools for the purpose of constructivist-based teaching in a school environment not only on theoretical grounds (Nicholson, 2018; Zhang et al., 2018), but also on an experimental basis.

In this study, we presented a tool suitable for measuring changes in semantic lexicon, the word association test, which can even be used in a classroom environment. For 7th graders, only three call words were selected for this purpose, since we observed a high dropout rate in this age group after the third keyword in our previous research. (Kovács & Murányi, 2024) For 9th graders, the number of word pairs was increased to 5, trusting that they would be able to complete the task. There are several options for the evaluation process, including calculating

the Garskof-Houston coefficient after manually processing the responses. The calculations can be performed with a validated LibreOffice macro, and the statistical analysis with free software. Another option is to use a LLM to evaluate responses by asking it to generate a mind map based on the call words. However, the reliability of this method could not be proven in this study, although it appears to be a very promising technique.

However, there are some limitations of this study. Based on the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995), t-tests may lead to false positive results in case of mixed wastewater and precipitation-liquid call word pairs (adjusted p values were 0.0725 and 0.11 respectively). However, the other results remained significant even after correction. Besides, The LLM used in this study was a smaller, weaker model with higher temperature; further research is planned with a stronger model and stricter settings after the ethical issues have been clarified.

Acknowledgements

I would like to thank the teachers for their work. I appreciate that they helped my research by guiding the game and having the tests written.

Ethics statement

This study received Eszterházy Károly Catholic University ethics approval (reference number: RK/144/2025). All methods were carried out in accordance with relevant guidelines and regulations.

Funding statement

The author received no financial support for the research, authorship, and/or publication of this article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bélanger, P. (2011). *Theories in adult learning and education*. Verlag Barbara Budrich. https://doi.org/10.3224/86649362

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cardellini, L. (2008). A note on the calculation of the Garskof-Houston relatedness coefficient/Una nota sobre el cálculo del coeficiente de relaciones de Garskof-Houston. *Journal of Science Education*, *9*(1), 48–51.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, *21*(1), 13–19. https://doi.org/10.1016/S0193-3973(99)00046-5

Chang, H.-Y. H. (2019). Escaping the gap: Escape rooms as an environmental education tool. *University of California*. https://nature.berkeley.edu/classes/es196/projects/2019final/ChangH_2019.pdf

Charkova, D. A. (2024). Utilizing gamification to promote pro-sustainable behavior among information technology students. *Discover Education*, *3*(1), 21. https://doi.org/10.1007/s44217-024-00105-x

Daru, K., & Tóth, Z. (2014). Óvodások időjárással kapcsolatos szóasszociációinak elemzése. In E. Juhász & T. Kozma (Eds.), *Oktatáskutatás határon innen és túl* (pp. 39-57.). Belvedere Meridionale. https://hera.org.hu/wp-content/uploads/2015/05/HuCER-2013-kotet.pdf

Garskof, B. E., & Houston, J. P. (1963). Measurement of verbal relatedness: An idiographic approach. *Psychological Review*, *70*(3), 277–288.

Gósy, M. (2005). *Pszicholingvisztika*. Osiris.

Hallinger, P., Wang, R., Chatpinyakoop, C., Nguyen, V.-T., & Nguyen, U.-P. (2020). A bibliometric review of research on simulations and serious games used in educating for sustainability, 1997–2019. *Journal of Cleaner Production*, *256*, 120358. https://doi.org/10.1016/j.jclepro.2020.120358

Iyamuremye, A., Niyonzima, F. N., Mukiza, J., Twagilimana, I., Nyirahabimana, P., Nsengimana, T., Habiyaremye, J. D., Habimana, O., & Nsabayezu, E. (2024). Utilization of artificial intelligence and machine learning in chemistry education: A critical review. *Discover Education*, *3*(1), 95. https://doi.org/10.1007/s44217-024-00197-5

Kapp, K. M. (2012). *The gamification of learning and instruction: Game-based methods and strategies for training and education*. Pfeiffer.

Kovács, M. (2024). Makró fejlesztése szóasszociációs teszt Garskof-Houston kapcsolati együtthatóval való kiértékelésére. *PedActa*, *14*(2), 25–31. https://doi.org/10.24193/PedActa.14.2.3

Kovács, M., & Murányi, Z. (2024). Rendhagyó kémia órák hatása 7. Osztályos diákok kognitív strukturájára. *Pedagogical Sections. Conference Proceedings*, 37–45. https://doi.org/10.36007/4966.2024.37

Mohebi, L. (2024). Empowering learners with ChatGPT: Insights from a systematic literature exploration. *Discover Education*, *3*(1), 36. https://doi.org/10.1007/s44217-024-00120-y

Muenchen, R. A. (2023). *BlueSky Statistics 10 User Guide*.

Nahalka, I. (2002). *Hogyan alakul ki a tudás a gyerekekben? Konstruktivizmus és pedagógia*. Nemzeti tankönyvkiadó.

Nicholson, S. (2015). *Peeking behind the locked door: A survey of escape room facilities*. https://scottnicholson.com/pubs/erfacwhite.pdf

Nicholson, S. (2018). Creating engaging escape rooms for the classroom. *Childhood Education*, *94*(1), 44–49. https://doi.org/10.1080/00094056.2018.1420363

Piaget, J. (1964). Development and learning. *Journal of Research in Science Teaching*, *2*(3), 176–186. https://doi.org/10.1002/tea.3660020306

Robertson, A. (1994). Toward constructivist research in environmental education. *The Journal of Environmental Education*, *25*(2), 21–31. https://doi.org/10.1080/00958964.1994.9941948

Robottom, I. (2004). Constructivism in environmental education: Beyond conceptual change theory. *Australian Journal of Environmental Education*, *20*(2), 93–101. https://doi.org/10.1017/S0814062600002238

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology*, *63*(3), 225.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research*, *81*(1), 4–28. https://doi.org/10.3102/0034654310393361

Tóth, Z. (2024). Examining the cognitive structure of elementary school students regarding science, energy sources, and health using the word association method. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(7), em2479. https://doi.org/10.29333/ejmste/14763

Zhang, X. C., Lee, H., Rodriguez, C., Rudner, J., Chan, T. M., & Papanagnou, D. (2018). Trapped as a group, escape as a team: Applying gamification to incorporate team-building skills through an ‘escape room’ experience. *Cureus*. https://doi.org/10.7759/cureus.2256