Constructing a Norm for Children's Scientific Drawing: Distribution Features Based on Semantic Similarity of Large Language Models

Yi ZHANG, Fan WEI, Yan WANG, Yanyan YU, Jianli CHEN, Zipo CAI, Xinyu LIU, Wei WANG, Peng WANG, Jingyi LI, Zhong WANG*

1. First Primary School, Fengtai, Beijing;
2. Fangzhuang Primary School, Fengtai, Beijing;
3. Hepingli No.9 Primary School, Dongcheng, Beijing;
4. College of Elementary Education, Capital Normal University, Beijing;
5. *Beijing Doers Education Consulting Co., Ltd, Beijing

Abstract: Using children's drawings to understand children's concept learning has been proven to be an effective method, but there are still two major problems in previous research: 1. The drawings heavily relies on the task, so the ecological validity of the conclusions is low; 2. The subjective interpretation of drawings is inevitable. To address these problems, this study uses the Large Language Model (LLM) to identify the drawing contents of 1096 children's scientific drawings (covering 6 scientific concepts), and uses the word2vec algorithm to calculate their semantic similarity. The study explores whether there are consistent drawing representations among children with the same theme, and attempts to establish a norm for children's scientific drawings, providing a baseline reference for following children's drawing research. The results showed that there were significant differences in the consistency of children's representations of different concepts, and there was a possibility of consistency bias, that is, the appearance of consistency representations misled LLM. At the same time, linear regression tests were used to analyze the relevant factors that affect children's representation. The results show that sample size and teaching strategies can affect the accuracy of LLM's image recognition, while the degree of conceptual abstraction may affect the consistency of representation.

Key words: Children's drawings, Norm, LLM, Semantic similarity, Distribution

1. Introduction

Children's drawing is a commonly used way of examining children's cognitive and emotional status [1]. The reason for this is not only because drawing is a pleasurable activity for children [2], but also because its effectiveness in examining children's psychological activities is considered reliable and has been extensively studied in many fields. Like from the perspective of individual learning, drawing is used to examine children's cognition of a certain scientific concept [3]. For example, Brechet et al. had over 250 French children draw their brains to understand how the "black box" of the brain works, while Marengo et al. analyzed over 600 drawings of Italian children to understand this problem [4].

However, there are still some unresolved problems in the psychological analysis of children's drawings, mainly reflected in two aspects: firstly, the problem of ecological validity. As drawing is a non-verbal representation, it must be digitized through effective encoding in order to be analyzed using statistical methods. Therefore, encoding method is currently the mainstream way to study children's drawing. For example, Tolsberg et al. used coding to understand the longitudinal development of math word problems among Estonian elementary school students [5]. So, these studies inevitably involve the problem of developing coding rules. This problem is particularly prominent in individual learning, as there are numerous scales or tools available in social psychology. Some drawing studies on this problem can use these general tools [6] [7]. However, for individual learning problems, researchers are forced to adopt self-developed coding rules. In a research topic on children's drawing under the journal “Frontiers in Psychology”, 71.4% of the studies used self-developed coding [8]. The high proportion naturally raises a question: how can these highly task dependent coding rules be transferred and generalized to other drawing studies? Obviously, existing drawing research lacks a unified norm. If this norm exists, different drawing tasks can be compared with it, then there is a common reference frame for cross task drawing research.

The second problem is more understandable: the understanding of drawing is one of the most prominent areas of individual heterogeneity, and different people's perspectives on the same drawing may be completely different [9]. Therefore, for researchers who currently do not utilize machines, consistency testing is an essential step - the same drawing must be identified by at least two researchers, and the results of these two identifications must pass consistency testing (such as Cohen Kappa coefficient [3] [5]). On the contrary, if machine learning models are used to identify drawings, this process may be omitted. Obviously, based on the same algorithm, machine learning may still have higher reliability than manual discrimination, even if it ignores consistency tests.

Based on the above two problems, this study aims to construct a norm for children's understanding of scientific concepts, and to use the Large Language Model (LLM) to understand and identify children's drawings, attempting to provide a reference frame for future research on children's conceptual understanding of drawing. This study focuses on three questions: 1. Do children have a common preference for drawing representation for the same drawing theme? 2. Is there a relationship between common preferences (if any) and LLM's ability to accurately identify drawing content? 3. What factors, if any, are related to common preferences?

2. Methods

2.1 Participants

Selected a total of 1207 students from two primary schools in Beijing. Most of them are in grades 5 and 6. The male to female ratio is 1.02:1. The textbooks studied involve two versions [10][11], and both of which are nationally approved textbooks.

We excluded drawings that were unclear, incomplete, or clearly did not belong to the prescribed content, as well as drawings with inconsistent LLM recognition results (see section 2.2.3 of this article), and ultimately included a total of 1096 drawings in the statistical analysis sample.

2.2 Procedure

2.2.1 Collection of the drawings

Firstly, we won't disrupt the order of teaching just for the sake of experiments, but required children draw concepts which they just learned. Secondly, to ensure the consistency and interpretability of the data, all teachers in this study are required to strictly follow the textbook to prevent teaching strategies and methods from interfering with the results. Thirdly, the prompt is: "We have just learned..., please use pictures to represent it”. Other than that, no hints will be given to prevent any misleading. Fourth, ensure that all students have completed their drawings.

2.2.2 Pretreatment of the drawings

Firstly, number and take photos of all the drawings. Secondly, preprocessing the photos includes two parts: one, removing students' personal privacy information such as names and student IDs; Another is to remove the theme (some students may unconsciously write the theme of the drawing on the drawing), in order to prevent prompting LLM. The solution is to use the "rubber stamp" function of Photoshop software to cover these contents with surrounding images.

2.2.3 Identified the drawings by LLM

The basic assumption of using LLM for image recognition in this study is that the angle and reason of LLM for image recognition are similar to the average human image recognition, so the results of image recognition can reflect the average level of image recognition of most people. This has also been confirmed by relevant research [12].

Firstly, input the photos of all the drawings into LLM for identification. The prompt is "What is depicted in this drawing? How did you see it from the drawing?", and copy and paste all the results and reasons generated by LLM into an Excel file.

In addition, to ensure the stability of LLM recognition, all drawings are independently recognized twice by LLM. Drawings with significantly different recognition results (such as the first recognition as a solar eclipse and the second recognition as a lunar phase) will be removed. The so-called 'independence' refers to deleting the conversation after one recognition is completed and re recognizing it. Since LLM does not capture conversations with front-end users and place them in its pre training set, the above operation method can ensure the independence of the two recognitions.

This study used ChatGLM-6.0 to identify drawings. The reason for adopting this LLM is that its image recognition models mainly use convolutional neural networks (CNN) and recurrent neural networks (RNN) (derived from the self-introduction of this LLM), and these two models have been proven to be efficient and reliable in image recognition (there are numerous related studies, so this article will not repeat the literatures here). And there have been some studies published using this LLM for image recognition [13] [14], further proving its reliability in image recognition. And its language generation method adopts a model based on the Transformer architecture, which has high efficiency and reliability in speech generation. Considering that there are many related studies, this article also does not cite references here for refer.

2.3 Semantic similarity analysis

2.3.1 Algorithm

This study uses the Word2vec algorithm to generate word vectors, and then uses the Cosine similarity algorithm to calculate semantic similarity values.

2.3.2 Analysis of distribution

Using heat maps to represent the semantic similarity between every two sentences in the semantic similarity matrix, and use a histogram to represent the specific distribution of similarity in [0,1].

Specifically, we calculated the ratio of the range of 75% values in the semantic similarity matrix of a certain concept to the overall range of values (hereinafter referred to as the "75% value concentration area"). This value indicates the extent to which the 75% semantic similarity values are concentrated (the smaller the value, the more concentrated it is), in order to examine the degree of concentration of the semantic similarity values represented by the concept. It is used to explore whether children have relatively consistent drawing representations for the same concept

2.4 Exploration of Related Factors

Finally, we investigated the effects of four factors, namely "accuracy", "sample size", "conceptual abstraction", and "teaching strategy", on the "accuracy", "semantic similarity mean", and "75% numerical concentration area" of children's drawing representations. The specific method is to use the above four factors as independent variables and the three represented factors as dependent variables, and conduct linear regression tests to detect correlation.

3. Results

3.1 General results

The total number of samples is 1096, covering 6 scientific concepts or process knowledges: Circuit, Solar eclipse, Boiling, Solar system, Increasing the carrying capacity of a ship, Life history of a plant. The average identification accuracy of LLM is 44.6%, and the average semantic similarity is 0.661. In addition, the mean of the "75% numerical concentration area" is 0.034. The data is detailed in Table 1.

From the overall situation, LLM has a low identification accuracy for image drawing content, but there is a certain degree of consistency in semantic similarity, and the result concentration is high. There is a certain degree of consistency preference in the overall representation, but this preference may lead to "collective misleading" (i.e., most children's drawing representations deviate significantly from the drawing content, and this deviation is highly consistent, thus misleading LLM).

Concept	Sample size	accuracy	Semantic similarity (average)	75%
Circuit	267	0.843	0.996	0.0002
Solar eclipse	182	0.44	0.995	0.0004
Boiling	48	0	0.975	0.001
Solar system	201	0.65	0.5	0.09
Increasing the carrying capacity of a ship	30	0.033	0	0
Life history of a plant	368	0.712	0.5	0.11

Table 1 Overall situation (where 75% refers to the 75% numerical concentration area, with smaller values being more concentrated)

3.2 Results of each scientific concepts

3.2.1 Circuit

This concept is mainly about lighting up a light bulb to let students know that an electrical circuit must be closed in order to function (i.e. forming a closed loop).

The average accuracy of LLM image recognition was 84.3%, with the highest accuracy among all six concepts, indicating that LLM can accurately understand the drawing theme through the representation methods of most students.

The average semantic similarity is 0.996, and 75% of the numerical values are concentrated within the 0.02% range of the theoretical value range ([0.9998, 0.9999]), indicating that the vast majority of students have a high degree of consistency in the morphological representation of circuits. For specific details, please refer to the upper left corner of Figure 1 and the upper left corner of Figure 2. The upper left corner of Figure 1 are the hot map and distribution, while the left corner of Figure 2 represents a representative drawing (i.e., a drawing with high semantic similarity within the 75% numerical concentration area, the same applies below).

3.2.2 Solar eclipse

This concept is mainly focusing on understanding the phenomenon and causes of solar eclipses, but does not require the use of light path diagrams.

The average accuracy of LLM image recognition is 44%, which is basically at the average level of six concepts.

The average semantic similarity is 0.975, and 75% of the numerical values are concentrated within the 0.1% range of the theoretical value range ([0.9977, 0.9987]), indicating that the vast majority of students have a high degree of consistency in the morphological representation of solar eclipses. However, it is worth noting that the accuracy of LLM image recognition is not high, indicating the possibility of consistency bias in students' representation of "solar eclipse". Refer to the upper right in Figure 1 and the middle of upper row in Figure 2 for details.

3.2.3 Boiling

This concept is mainly about boil water through heating. The textbook requires students to know the temperature of boiling and the process of changes of water.

The average accuracy of LLM image recognition is 0, ranking at the bottom of the six concepts, indicating that all students' representations have misled LLM.

The average semantic similarity is 0.995, and 75% of the numerical values are concentrated within the 0.04% range of the theoretical value range ([0.9995, 0.9999]), indicating that the vast majority of students have a high degree of consistency in the morphological representation of water evaporation. However, LLM's image recognition is completely incorrect, indicating that all students' representations are incorrect. This result is very surprising, as can be seen from the representative drawing in the top right of Figure 2 where the problem lies: all the drawings depicted were heat water by alcohol lamps. This experiment was actually intended to heat water to make it boil, not boiling itself, but the students' focus was entirely on the experiment itself - rather than what the experiment’s meaning, which is truly surprising. See left in Figure 1 and top right in Figure 2 for details.

3.2.4 Solar system

This concept is mainly about the solar system, especially the Sun, 8 planets and their relative positions, but does not require their motion state and trajectory.

The average accuracy of LLM image recognition is 0.65, which is higher than the mean, indicating that most students can correctly represent this concept.

The average semantic similarity is 0.5, and 75% of the numerical values are concentrated within the 9% range of the theoretical value range ([0.46, 0.55]). Although the semantic similarity is moderate and the similarity values are relatively concentrated, 9% is already ranked second to last in the concentration of the six concepts, indicating that students' representations of this concept are not completely consistent and there are certain differences. Refer to Figure 1 on the right and Figure 2 on the bottom left for details.

3.2.5 Increasing the carrying capacity of a ship

This concept is mainly about the close relationship between buoyancy and the volume of discharged water by changing the volume and load capacity of the ship. But the presentation of this lesson is quite unique: first, use aluminum foil of the same area to make boats with different bottom areas, and then add iron gaskets to different boats to increase their weight and test their load-bearing capacity. Ultimately, guide students to identify factors associated with buoyancy. In other words, the experiment in this class is actually divided into two steps: first, make the ship and change its volume; secondly, increase the load capacity.

Surprisingly, the average semantic similarity is 0, and the 75% numerical concentration area is also 0. This indicates that LLM's recognition of all drawings related to this concept is completely unrelated to each other. Since they are not related to each other, we not display representative drawings (as there are no representative drawings) in this concept. Refer to the bottom left of Figure 1 for details.

3.2.6 Life history of a plant

This concept is mainly about the life history of plants, that is, the entire process of Impatiens from seed to flowering and fruiting. In the previous classes, students have already gone through cultivation, care and other work, and recorded the growth and changes of Impatiens. Therefore, the main purpose of this lesson is to establish an overall impression of the development and changes of Impatiens throughout its life by reviewing and summarizing previous recorded materials.

The average accuracy of LLM image recognition is 0.712, ranking second among the six concepts, indicating that most students are able to correctly represent this concept.

However, the results of semantic similarity are not concentrated, with an average semantic similarity of 0.5. 75% of the numerical values are concentrated within the 11% range of the theoretical value range ([0.45, 0.56]). This result is very similar to the concept of "solar system", indicating that students' representations of this concept are not completely consistent and there are certain differences. Please refer to the bottom right of Figure 1 and the middle and right rows of Figure 2 for details.

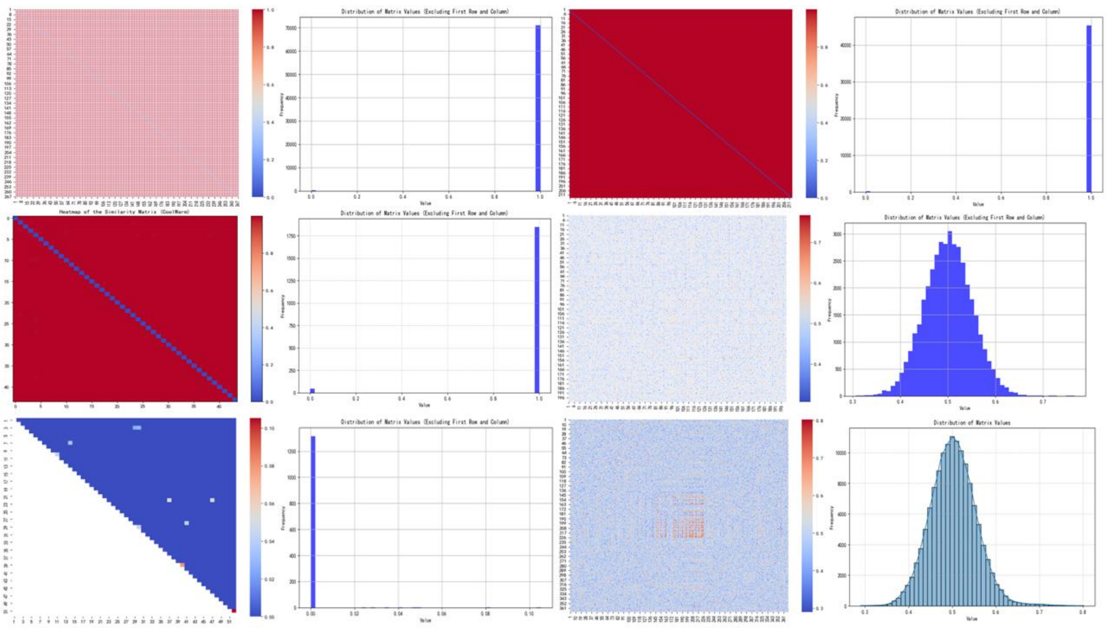


Figure 1: Semantic similarity heatmap and distribution map of 6 concepts

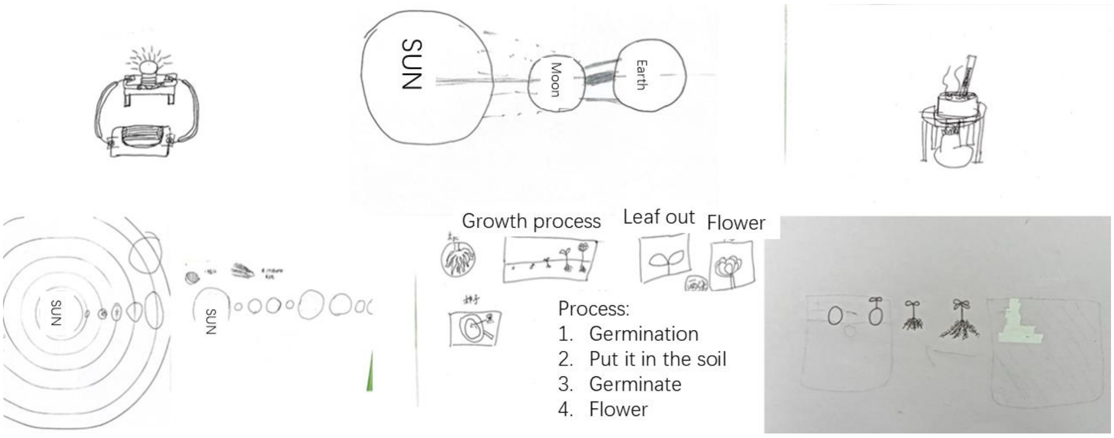


Figure 2: Representative drawings of 5 concepts (excluding "Increasing the carrying capacity of a ship")

3.3 Analysis of Related Factors

3.3.1 Word Vector Dimension

What reasons may these results be related to? We first analyzed the factors of word vector dimension. We calculated the word vector dimensions of six concepts using the TF-IDF algorithm, as shown in Table 2. Then, with "accuracy", "semantic similarity mean", and "75% numerical concentration area" as dependent variables, and word vector dimension as independent variables, linear regression models were established separately. The results showed that the coefficient of determination (R^2) of the three variables were 0.49, 0.001, and 0.46, respectively (as shown in the upper left row of Figure 3), indicating that the word vector dimension cannot well explain the correlation between accuracy and semantic similarity. That is to say, the above results are not sensitive to the complexity of LLM generated statements.

3.3.2 Relationship with sample size

Is there a relationship between the above results and the total sample size? Using the method described in 3.3.1, we establish a linear regression model with "sample size" as the independent variable. The results showed that the determination coefficients (R^2) of the three were 0.83, 0.04, and 0.43, respectively (as shown in the top three right row of Figure 3), indicating that sample size has a high explanatory power for accuracy, but it is not closely related to the other two factors.

3.3.3 Relationship with the level of conceptual abstraction

Based on Piaget's theory of developmental stages in children, the abstract thinking ability of elementary school students has not yet been fully formed [15]. Therefore, we speculate whether the above results are related to the degree of abstraction of concepts? That is to say, the higher the level of abstraction, the lower the accuracy and the lower the semantic similarity?

For this purpose, we encode the abstraction level of all six concepts. The coding rule is: the research object or phenomenon that appears in the textbook with a unique representation is 1, and the representation that is not unique in the textbook is 2. Based on this rule, the following encoding is obtained:

Concepts	Codes	Reasons of encoded
• Circuit	1	Unique form, a closed loop is like a 'circle'
• Solar eclipse	1	Require to remember the appearance of the first loss, severe food, and other moments, and the textbook provides photos
• Boiling	1	The only research object is water
• Solar system	1	The textbook provides the standard appearance of the solar system
• Increasing the carrying capacity of a ship	1	The textbook strictly specifies the types of ships and provides photos after production
• Life history of a plant	2	Textbooks do not have clear representations of life history and appear in the form of record sheets

Table 2: Abstraction level encoding and specific reasons for 6 concepts

As mentioned above, we still use "abstraction degree" as the independent variable and "accuracy", "semantic similarity mean", and "75% numerical concentration area" as dependent variables to construct linear regression functions. The results show that the coefficient of determination (R^2) of the three variables are 0.13, 0.04, and 0.52, respectively (the three images in the bottom left row of Figure 3). It can be inferred that the degree of abstraction of a concept is not closely related to "accuracy" and "semantic similarity mean", but may be related to the "75% numerical concentration area" (the degree of concentration of semantic similarity).

3.3.4 Relationship with Teaching Strategies

This is what we discovered after carefully examining the drawings: we found that the concept of "Boiling" originally required students to understand the process of water to boiling, but most of them drew the experimental process (namely the appearance of the experimental setup). This experiment was originally intended to demonstrate the process of water change, but it was used by students as a representation of boiling. In other words, the experiment had a counterproductive effect here.

Therefore, we encoded the role of experiments in teaching and used it as an independent variable to conduct linear regression tests on the above three factors. The coding rule is as follows: if there is no experiment or the experimental result is the teaching result, it is considered as 1; The experimental results are not teaching results, but only prove that the teaching results are 2 (simple drawing and activities are not considered experiments). The encoding result is as follows:

Concepts	Codes	Reasons of encoded
• Circuit	1	The experimental result is to form a closed loop and light up the light bulb
• Solar eclipse	1	There are no hands-on experiments in this class
• Boiling	2	The experimental results only proved where the water went, not the evaporation itself
• Solar system	1	There are no hands-on experiments in this class
• Increasing the carrying capacity of a ship	2	Contains two experiments, the first experiment is not related to the teaching results, but only verifies whether the ship can float
• Life history of a plant	1	There is no experiment in this class, only the planting activity of Impatiens balsamina

Table 3: Teaching strategies encodings for 6 concepts

The coefficient of determination (R^2) for modeling are 0.87, 0.11, and 0.24, respectively (as shown in the bottom right of Figure 3), indicating that the consistency between experimental results and teaching results can effectively explain the accuracy. Specifically, it can effectively solve the problem of "collective bias" in students' drawings: students are more concerned about the experiment itself, but if the experiment deviates from the teaching objectives, they are likely to choose to remember the experiment rather than the real objectives.

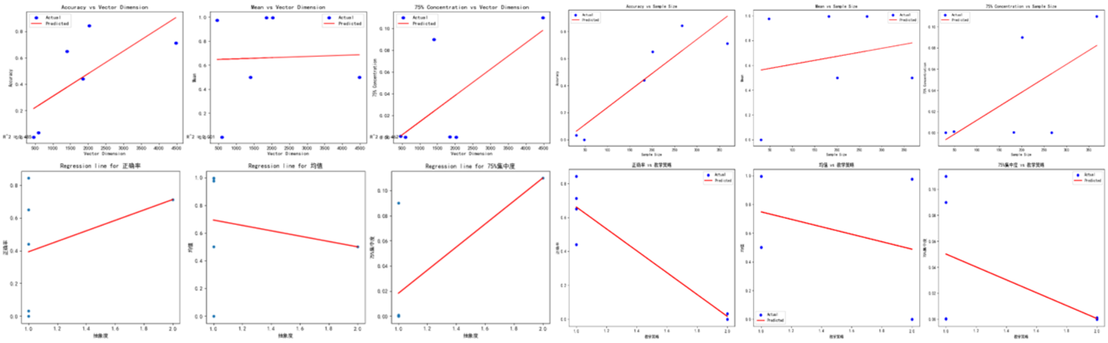


Figure 3: Correlation analysis of four factors

4. Disccussion

This article outlines the general appearance of six scientific concepts in children's eyes through 1096 children's scientific drawings, and explores whether there is consistency representation in children's scientific drawings through LLM image recognition and semantic similarity analysis. The results show that there are significant differences between different concepts, with some concepts having consistent representations while others do not. At the same time, it was found that the consistency of the representation is independent of the accuracy, indicating the existence of consistency bias.

In the subsequent exploration of influencing factors, we used linear regression to investigate the effects of "word vector dimension", "sample size", "abstraction", and "teaching strategy" on drawing. It was found that accuracy is the most sensitive indicator, and both "sample size" and "teaching strategy" can explain it relatively well. And the most intriguing aspect of this is the teaching strategy, as the results of regression analysis seem to conflict to some extent with the ideas we advocate now.

The 2022 version of the "National Science Education Standards" of China particularly emphasizes the role of thinking in science classrooms. Especially reflected in the relationship between thinking argumentation and exploratory practice: "Based on evidence and logic... establish the relationship between evidence and explanation" "analyze evidence and draw conclusions, interpret and evaluate the results" [16]. This exposition positions the exploration experiment at the lower level of reasoning and argumentation, that is, classroom experiments serve reasoning and argumentation rather than directly serving teaching results (knowledge objectives). In other words, the experiments advocated by the standards should not only focus on knowledge objectives, but may require further reasoning and abstraction in order to derive the true teaching objectives.

However, a large amount of evidence has shown that children may find it difficult to meet this requirement. In fact, this result shifts the focus to another question: Which is more competitively in children's cognition about visual processing or semantic processing? Obviously, experimental phenomena rely more on visual processing, and reasoning and proving conclusions through experiments require a significant amount of semantic processing. In this regard, the results clearly support more visual processing. Many studies have confirmed that more complex semantic processing can only occur in older children [17] [18]. In summary, these pieces of evidence explain why children in this study tend to mistake experimental results for teaching outcomes. And this result also reveals that when setting teaching objectives, we should not only rely on curriculum concepts, but also pay attention to the cognitive development laws of children.

Reference

[1] Fabris MA, Lange-Küttner C, Shiakou M and Longobardi C (2023) Editorial: Children's drawings: evidence-based research and practice. *Front. Psychol.* 14:1250556. https://doi.org/10.3389/fpsyg.2023.1250556

[2] P. Bozzato,Matteo Angelo Fabris &C. Longobardi. Gender, stereotypes and grade level in the draw-a-scientist test in Italian schoolchildren. *International Journal of Science Education.* 2021.10, 2640-2662, https://doi.org/10.1080/09500693.2021.1982062

[3] Brechet C, Blanc N, Mortier A and Rossi S (2022) Draw me a brain: The use of drawing as a tool to examine children's developing knowledge about the “black box” . *Front. Psychol.* 13:951784. https://doi.org/10.3389/fpsyg.2022.951784

[4] Marengo D, Settanni M, Longobardi C and Fabris MA (2022) The representation of bullying in Italian primary school children: A mixed-method study comparing drawing and interview data and their association with self-report involvement in bullying events. *Front. Psychol.* 13:862711. https://doi.org/10.3389/fpsyg.2022.862711

[5] Tolsberg K, Põldre S and Kikas E (2022) Learner-generated drawing as a learning strategy. The effect of teacher-guided intervention program “Learning with Understanding” on composing drawings in math word problems in the primary grades. *Front. Educ.* 7:962067. https://doi.org/10.3389/feduc.2022.962067

[6] Restrepo Cervantes, D., Chamorro Coneo, A., Bolivar Pimiento, D. et al. A Psychometric Analysis of the Moral Disengagement Scale (MDS) in Association to Bullying Roles in Colombian Youth. *Int Journal of Bullying Prevention* (2024). https://doi.org/10.1007/s42380-024-00215-y

[7] Marengo D, Settanni M, Longobardi C and Fabris MA (2022) The

representation of bullying in Italian primary school children: A mixed-method study comparing drawing and interview data and their association with self-report involvement in bullying events. *Front. Psychol.* 13:862711. https://doi.org/10.3389/fpsyg.2022.862711

[8] Front. Psychol. Children's Drawings: Evidence-based Research and Practice. https://www.frontiersin.org/research-topics/24933/childrens-drawings-evidence-based-research-and-practice/overview

[9] Beltzung B, Pelé M, Renoult JP and Sueur C (2023) Deep learning for studying drawing behavior: A review. *Front. Psychol.* 14:992541. https://doi.org/10.3389/fpsyg.2023.992541

[10] Bo YU. 2017. *Science.* Beijing: Education Science Press

[11] Wei DUAN, Xiang PENG. 2017. *Science*. Changsha: Hunan Science and Technology Press

[12] YA Cheng et al. A neural geometry theory comprehensively explains apparently conflicting models of visual perceptual learning. biorxiv. https://doi.org/10.1101/2023.11.13.566963

[13] Maolin Zhang, Xianyong Li, Yajun Du et al. A Multimodal Large Language Model for Forestry in pest and disease recognition. *Research Square.* https://doi.org/10.21203/rs.3.rs-3444472/v1

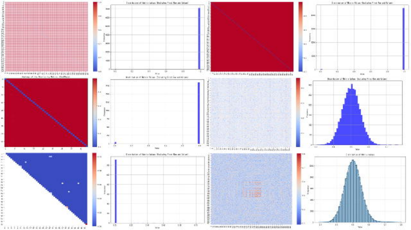
[14] Shuo Yang, Zirui Shang, Yongqi Wang et al. Data-free Multi-label Image Recognition via LLM-powered Prompt Tuning. *Arxiv.* https://doi.org/10.48550/arXiv.2403.01209

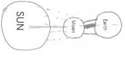
[15] Galotti KM. 2017. *Cognitive Psychology: In and Out of the Laboratory*. Beijing: China Machine Press, 218

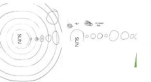
[16] Ministry of Education, PRC. *National Science Education Standard of Compulsory Education*. 2022. Beijing: Beijing Normal University Press

[17] Tai-Li Chou, James Booth, Tali Bitan et al. Developmental and Skill Effects on the Neural Correlates of Semantic Processing to Visually Presented Words[J]. *Human Brain Mapping.* 27(11), 915-924, https://doi.org/10.1002/hbm.20231

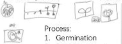
[18] Lucy A. Henry henrylc@lsbu.ac.uk, David Messer. Phonological, visual, and semantic coding strategies and children's short-term picture memory span[J]. Quarterly Journal of Experimental Psychology. 2012.10, https://doi.org/10.1080/17470218.2012.672997







Growth process Leaf out Flower



Process:
1. Germination
2. Put it in the soil
3. Germinate
4. Flower