

Using Large Language Models To Diagnose Math Problem-solving Skills At Scale

Hyoungwook Jin*

jinhw@kaist.ac.kr

School of Computing, KAIST
Daejeon, Republic of Korea

Yoonsu Kim*

yoonsu16@kaist.ac.kr

School of Computing, KAIST
Daejeon, Republic of Korea

Yeon Su Park

yeonsupark@kaist.ac.kr

School of Computing, KAIST
Daejeon, Republic of Korea

Bekzat Tilekbay

tilekbay@kaist.ac.kr

School of Computing, KAIST
Daejeon, Republic of Korea

Jinho Son

sjhfam@algorithmllabs.co.kr

Algorithm LABS
Seoul, Republic of Korea

Juho Kim

juhokim@kaist.ac.kr

School of Computing, KAIST
Daejeon, Republic of Korea

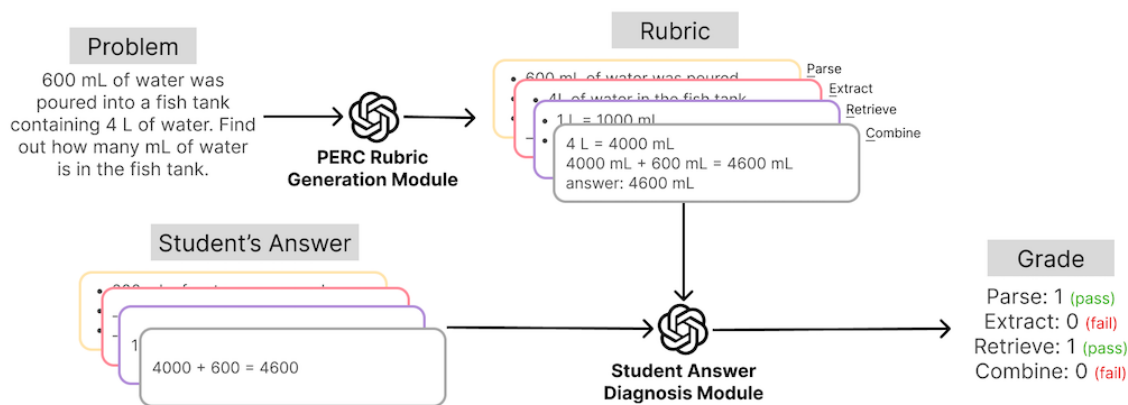


Figure 1: An overview of an LLM-based pipeline for diagnosing students' answers to math word problems. Our pipeline first generates a rubric and compares it to students' answers to determine their mastery.

ABSTRACT

Personalized feedback, tailored to students' needs and prior knowledge, is essential for fostering mathematical problem-solving skills. However, personalized feedback is often limited to one-to-one tutoring or small classrooms as it requires instructors' in-depth diagnosis of cognitive processes employed in students' answers. We propose a large language model (LLM) pipeline that diagnoses students' problem-solving skills from their answers at scale in elementary school math word problems. Based on prior literature and an interview with a math education expert, we developed PERC, a framework composed of four problem-solving stages that students can follow: Parse, Extract, Retrieve, and Combine. The framework facilitates diagnosis by externalizing students' step-by-step problem-solving processes and allowing our pipeline to analyze each stage

individually. Our LLM pipeline diagnoses each stage by (1) generating rubrics and (2) comparing students' answers with the rubrics. We fine-tuned our LLM pipeline with 71 math problem-rubric pairs and 128 problem-answer-grade triplets collected from elementary school students. We evaluated our pipeline's diagnosis accuracy against vanilla GPT-3.5 and vanilla GPT-4 with automatic and expert evaluations. The results showed the potential of our approach in improving the end-to-end diagnosis accuracy of LLMs, and expert evaluation provided specific aspects that should be improved.

CCS CONCEPTS

• Applied computing → Education.

KEYWORDS

Mathematical problem-solving skills, Educational diagnosis at scale, Large language models

ACM Reference Format:

Hyoungwook Jin, Yoonsu Kim, Yeon Su Park, Bekzat Tilekbay, Jinho Son, and Juho Kim. 2024. Using Large Language Models To Diagnose Math Problem-solving Skills At Scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*, July 18–20, 2024, Atlanta, GA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3657604.3664697>

*Both authors contributed equally to this research.

1 INTRODUCTION

Although online-based learning (e.g., MOOCs) has significantly improved accessibility to learning mathematical problem-solving [1], the uniform guidance and feedback provided by online platforms and technologies limit the effectiveness of learning (e.g., showing only solutions without explanations adaptive to students' answers) [14]. Compared to these one-size-fits-all scaffolds, feedback personalized to students' different prior knowledge can elicit the best learning effects and experiences [11, 20].

However, it has been challenging to scale personalized feedback to online-based learning. Personalized feedback involves a fine-grained diagnosis of each student's skills and adaptive suggestions for improvement [11, 19]. The diagnosis requires in-depth comprehension of students' solutions written in natural language and mathematical expressions, making it heavily reliant on teachers' expertise and manual effort [6]. Consequently, the personalized diagnosis remains feasible only in small classroom settings.

We investigate the utility of large language models (LLMs) for diagnosis at scale. Recent work highlights the exceptional performance of LLMs in understanding and solving mathematical problems [7]. By utilizing such capabilities, we explore using LLMs to provide tailored diagnoses of students' problem-solving skills at scale. We anticipate that, through clear evaluation rubrics for diagnosis (i.e., a list of correct answers for a problem), LLM can leverage its natural language understanding and reasoning to evaluate students' answers [8]. In this project, we focus on diagnosing the problem-solving skills of elementary school math word problems.

We developed a problem-solving framework and an LLM pipeline to scaffold and diagnose four primary skills in mathematical problem-solving. First, we designed a PERC framework with parsing, extracting, retrieving, and combining stages based on prior literature on problem-solving processes [9, 13, 15] and an expert interview about elementary school education. The framework emphasizes the skills to digest textual information, which contemporary young students often struggle with. Then, we developed a two-stage LLM pipeline to provide personalized diagnoses for students' math problem-solving skills at scale. Using GPT-3.5, our pipeline first generates the PERC rubrics for the given math word problem and uses them to diagnose students' mathematical problem-solving competencies based on their answers. To further enhance the accuracy of the pipeline, we collected and fine-tuned GPT-3.5 with a dataset of 71 problems, 71 expert-generated rubrics for each stage, and 128 responses from elementary school students with expert diagnosis results.

To assess the efficacy of our LLM pipeline, we conducted technical and expert evaluations using vanilla GPT-3.5 and GPT-4 models as baselines. We observed that the rubric generation module effectively replicates mathematical expressions used by experts but is weak at including all problem-solving details compared to GPT-4. Experts remarked that our pipeline generates satisfactory rubrics and diagnoses, but the rubrics should avoid including the knowledge that students have not learned. We discuss the potential of using LLMs for math diagnosis at scale and future directions for improvement.

This paper makes the following contributions:

- PERC (Parse-Extract-Retrieve-Combine) framework that can scaffold elementary school students to tackle math word

problems and externalize their problem-solving process step-by-step.

- An LLM pipeline that can diagnose students' mathematical problem-solving skills based on the PERC framework.

2 FRAMEWORK DEVELOPMENT

We reviewed problem-solving stages studied in literature and interviewed a math education expert to identify important stages and skills in elementary school math problem-solving. We reorganized the insights and findings into the PERC framework.

2.1 Existing Problem-Solving Stages

One of the most commonly used frameworks proposed to incorporate comprehensive problem-solving methodologies in mathematics was suggested by Polya [13]. He identified four stages to enhance problem-solving strategy: (1) understanding the problem, (2) devising a plan, (3) carrying out the plan, and (4) looking back. This framework suggests devising a *plan* before the computations to approach problems thoughtfully and systematically.

Polya's framework has influenced many other researchers to propose new approaches that utilize the four-stage problem-solving strategy in various ways [4, 10, 16]. For example, Goulet et al. highlight the need to establish links between the explicit problem content and students' knowledge while devising a plan [3]. Schoenfeld concluded four important categories in solving problems: (1) resources, (2) heuristics, (3) control, and (4) belief systems [15]. Similarly, Krulik and Rudnick further specified the set of heuristics to be successful in all levels of problem-solving: (1) read, (2) explore, (3) select a strategy, (4) solve, and (5) look back and extend [9]. It was found that utilizing such approaches to give step-by-step guidelines in problem-solving significantly improves learning in mathematics [2, 16, 18], especially for students with a low level of achievement [10].

2.2 Domain Expert Interview

To understand how the previous frameworks can be applied in the real-world setting of elementary-level education, we conducted a semi-structured interview with a domain expert who has been working in mathematics education for 30 years. We asked questions about (1) the characteristics of elementary school students, (2) how Polya's framework can be improved, (3) how teachers evaluate students' mathematics skills, and (4) what to consider while designing the LLM to evaluate students' answers.

The expert expressed the need to adapt Polya's framework for elementary-level students. Besides, the expert noted that many elementary school students struggle to read the given problem correctly, thus, math education should also focus on *comprehension* before *plan* to proceed to computation. The expert emphasized the severity of this issue among the current generation of elementary school students who tend to consume media passively, such as videos. This finding aligns with previous studies that explored mathematical problem-solving according to Polya's four-stage framework [12, 23]. They found that elementary school students experienced difficulties in solving mathematical problems, especially in comprehension of the problem [23], which can cascade into errors in the following stages.

2.3 PERC Framework

Based on the literature review and interview findings, we propose a *Parse-Extract-Retrieve-Combine*, PERC framework for mathematics education of elementary school students. We reorganized existing frameworks into the PERC framework to focus on *comprehension* skills in problem-solving. This framework can serve as a lens to analyze the information digestion process in math and a template for collecting students' in-depth problem-solving processes at scale.

The PERC framework comprises four stages:

- **Parse:** Parse the problem in mathematical units to identify key information and list requirements.
- **Extract:** Extract the mathematical expression of each parsed unit of the problem to construct a concrete plan.
- **Retrieve:** Retrieve external mathematical information and formulas that are not present in the problem but are needed to solve the problem.
- **Combine:** Combine the internal information from *Extract* stage and external information from *Retrieve* stage to systematically solve the problem according to the plan.

The framework can guide elementary school students to approach problems step-by-step and externalize their problem-solving process. For example, students can be given questions corresponding to the four stages and structurally work through each stage to solve a problem. The example scenario with the PERC framework is illustrated in Figure 2.

3 PERC DATASET

While PERC can scaffold students to tackle problems step-by-step, we aim to push it further by diagnosing students' answers to each stage and giving personalized feedback with LLMs. To develop and evaluate our LLM pipeline, we constructed a PERC dataset based on 71 problem-rubric pairs and 128 problem-answer-grade triplets based on the PERC framework (Figure 2).

First, we collected a set of 71 mathematical problems with diverse difficulties concerning K3-6 public curriculum topics in South Korea (e.g., arithmetics, unit conversion). We asked two math educators, each with over 20 years of teaching experience, to collaboratively make the rubrics for the problems. The rubric described the ideal response for each PERC stage.

Then, we recruited 55 K3, 45 K4, 46 K5, and 39 K6 students and asked them to solve the problems based on the PERC framework on paper worksheets. Then, we manually transcribed the problem-answer pairs into \LaTeX format to facilitate data-feeding to LLMs. We excluded students' answers that were unrecognizable by humans. As a result, we compiled 128 problem-answer pairs.

With the collected 128 problem-answer pairs, we recruited 64 experienced math teachers to grade the answers. They compared students' answers with the rubrics and gave a score of 1 if the answer was correct or 0 otherwise for each PERC skill. Each answer was graded only by one teacher.

4 DIAGNOSING PERC AT SCALE

We introduce an LLM-based pipeline to automate the diagnosis of the PERC framework. We also report a rudimentary evaluation of our pipeline.

	Model	Parse	Extract	Retrieve	Combine
BLEU (Textual similarity)	Ours	0.59	0.12	0.00	0.23
	vanila gpt-3.5	0.42	0.06	0.00	0.16
	vanila gpt-4	0.47	0.03	0.00	0.10
Embedding similarity	Ours	0.09	0.25	0.38	0.12
	vanila gpt-3.5	0.16	0.28	0.45	0.16
	vanila gpt-4	0.13	0.28	0.42	0.17

Table 1: Evaluation Results of PERC Rubric Generation Module

Model	Parse	Extract	Retrieve	Combine
Ours	92.59	88.89	81.48	96.30
baseline (vanila gpt-3.5)	88.89	81.48	77.78	96.30
baseline (vanila gpt-4)	85.19	81.48	77.78	85.19

Table 2: Evaluation Results of Student Answer Diagnostic Module

4.1 LLM Pipeline

We implemented an LLM pipeline to diagnose student responses to the PERC framework (Figure. 1). The LLM pipeline consists of three components: 1) a PERC rubric generation module, 2) an answer diagnosis module, and 3) a feedback generation module. We chose a multi-staged design for the pipeline because the division of work is known to be effective in improving the accuracy and controllability of LLMs [22].

4.1.1 PERC Rubric Generation Module. This module receives a math word problem and generates the corresponding PERC rubric for grading. The module ensures scalability by automatically creating the PERC rubric for a given problem. By guiding LLMs to solve mathematical problems following the PERC stages, we maximize the inferential capabilities of LLMs, akin to effective prompting techniques such as chain of thought [21]. To develop this module, we gave explanations about PERC as instructions and fine-tuned the GPT-3.5-0125 model with 80% of the problem-rubric pairs in the PERC dataset.

4.1.2 Student Answer Diagnosis Module. This module receives a math word problem, the PERC rubric generated by the *PERC Rubric Generation Module*, and a student's answer. The module diagnoses the student's mathematical competency and provides binary judgments (i.e., 1 for pass and 0 for fail) for each PERC competency. We developed this module by fine-tuning the GPT-3.5-0125 model with 80% of the problem-answer-grade triplets.

4.1.3 Feedback Module. The last module provides feedback based on the diagnosis result. There can be 16 combinations of diagnosis results (i.e., 0 or 1 for each PERC). According to the combination, our pipeline gives feedback predefined by the math education expert.

4.2 Technical Evaluation

We conducted technical evaluations of the *PERC Rubric Generation Module* and the *Student Answer Diagnosis Module* by comparing them with the vanilla gpt-3.5-turbo-0125¹ and gpt-4-0613² with 4-shots examples as baseline with the remaining 20% of the dataset.

¹<https://platform.openai.com/docs/models/gpt-3-5-turbo>

²<https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>





Problem	Framework	Expert Rubric	Student Answer	Grading
600 mL of water was poured into a fish tank containing 4 L of water. Find out how many mL of water is in the fish tank.	 Parse	600 mL of water was poured into a fish tank containing 4 L of water. Find out how many mL of water is in the fish tank.	600 mL of water was poured into a fish tank containing 4 L of water. Find out how many mL of water is in the fish tank.	1
	 Extract	<ul style="list-style-type: none"> 4L of water in the fish tank 600 mL of water was poured. → Convert the amount of water to mL.	<ul style="list-style-type: none"> - Fish tank containing 4L of water - 600 mL of water was poured. 	0
	 Retrieve	<ul style="list-style-type: none"> 1 L = 1000 mL L is added to L, and mL is added to mL. 	1 L = 1000 mL	1
	 Combine	4 L = 4000 mL 4000 mL + 600 mL = 4600 mL answer: 4600 mL	4000 + 600 = 4600	0

Figure 2: An example data point of the PERC dataset. The dataset comprises 71 unique math word problems, expert-generated rubrics (i.e., exemplary answers) for each problem, 128 students’ answers, and experts’ grading for each PERC skill.

Model	Parse	Extract	Retrieve	Combine
Ours	45	33	27	28
baseline (vanila gpt-3.5)	38	34	26	29
baseline (vanila gpt-4)	42	38	26	39

Table 3: Expert Evaluation Results

4.2.1 *PERC Rubric Generation Module Evaluation.* We evaluated textual and semantic similarities between the module-generated rubrics and expert-generated ones for each PERC (Table 1). We looked into both similarities to reflect the preciseness in mathematical expressions and the flexibility of natural language in math problem-solving. We chose BLEU for the metric of textual similarity to count exact word matches, and OpenAI’s embedding model³ for semantic similarity to allow interchangeable words and expressions.

We found that the BLEU scores were high for **Parse**, reflecting the effectiveness in identifying key information from given math problems. In the **Retrieve** phase, although the textual similarity yielded a score of 0, the high embedding similarity suggests meaningful retrieval of external information necessary for problem-solving. Furthermore, our model exhibited higher textual similarity than the baseline, yet lower embedding similarity. This discrepancy implies that while our model effectively captured the textual representation of mathematical concepts, it struggled to capture their semantic meaning.

4.2.2 *Student Answer Diagnosis Module Evaluation.* We compared the accuracy of the model’s binary (0 or 1) diagnosis results for each PERC against the experts’ grades. The results, depicted in Table 2, demonstrate that our pipeline outperformed baseline models in accuracy according to expert evaluation, which indicates its effectiveness in evaluating mathematical competencies.

4.3 Expert Evaluation

For *PERC Rubric Generation Module*, we conducted qualitative expert evaluations on the generated rubrics. We asked a mathematics education expert to evaluate the PERC-specific rubrics generated by the module and those generated by vanilla GPT-3.5 and GPT-4. The evaluation was conducted blindly to ensure that the expert was unaware of the type of output being assessed. The rubrics were

scored on a scale of 1 to 3, with 1 indicating error or irrelevant content, 2 for correct but limited for an ideal rubric, and 3 for satisfactory rubrics. In Table 3, we report cumulative scores of the PERC rubrics generated by the three models for 15 math problems in the PERC dataset. The expert concluded that our model for **Parse** is satisfactory, but refinements are needed for **Extract**, **Retrieve**, and **Combine** stages to align more closely with the student levels and curriculum. Specifically, the expert suggested restricting these stages to the learning materials of the target grade level.

5 FUTURE WORK

We propose several directions for future work. First, we need to confirm the efficacy of the PERC framework in improving students’ math problem-solving skills. Acknowledging that a single interview is insufficient, we will conduct interviews with more domain experts. Furthermore, we will run empirical studies to verify whether the framework can improve students’ information processing ability compared to the answer templates from existing research [10]. We will conduct longitudinal comparative studies to measure students’ perceived efficacy, learning gain, and metacognition.

We will advance our LLM pipeline to improve the PERC rubric generation and achieve fine-grained diagnosis. As pointed out in the expert evaluation, we will improve the *PERC Rubric Generation Module* by preventing it from using knowledge beyond the curriculum level of problems. We will also advance the *Student Answer Diagnosis Module* to generate finer-grained scores (i.e., rational numbers instead of binary classification). This involves decomposing the knowledge units in the PERC rubrics, comparing them with student answers, and computing coverage. Fine-grained grading can improve the explainability of diagnosis and personalized feedback.

Lastly, we will publicly release our PERC dataset, consisting of math word problems, PERC rubrics, student answers, and expert grading. The dataset will serve as a benchmark for evaluating the accuracy of automated mathematical skill diagnosis [5], a training dataset for math-reasoning tasks [21], and real-world data for learning analytics [17].

ACKNOWLEDGMENTS

This work was supported by Algorithm LABS & Hamamath.

³<https://platform.openai.com/docs/guides/embeddings/embeddings>

REFERENCES

- [1] Vanessa Chang, Christian Gütl, and Martin Ebner. 2018. Trends and opportunities in online learning, MOOCs, and cloud-based tools. *Second handbook of information technology in primary and secondary education* (2018), 935–953.
- [2] Esen Ersoy and Belgin Bal-Incebacak. 2017. The evaluation of the problem solving in mathematics course according to student views. In *ITM Web of Conferences*, Vol. 13. EDP Sciences, 01012.
- [3] Marie-Pier Goulet-Lyle, Dominic Voyer, and Lieven Verschaffel. 2020. How does imposing a step-by-step solution method impact students' approach to mathematical word problem solving? *ZDM* 52, 1 (2020), 139–149.
- [4] Nourooz Hashemi, Mohd Salleh Abu, Hamidreza Kashefi, Mahani Mokhtar, and Khadijeh Rahimi. 2015. Designing learning strategy to improve undergraduate students' problem solving in derivatives and integrals: A conceptual framework. *Eurasia Journal of Mathematics, Science and Technology Education* 11, 2 (2015), 227–238.
- [5] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).
- [6] Jessica Hoth, Martina Döhrmann, Gabriele Kaiser, Andreas Busse, Johannes König, and Sigrid Blömeke. 2016. Diagnostic competence of primary school mathematics teachers during classroom situations. *ZDM* 48 (2016), 41–53.
- [7] Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398* (2023).
- [8] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2023. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. *arXiv:2309.13633 [cs.HC]*
- [9] Stephen Krulik and Jesse A Rudnick. 1988. *Problem Solving: A Handbook for Elementary School Teachers*. ERIC.
- [10] Chien I Lee. 2016. An appropriate prompts system based on the Polya method for mathematical problem-solving. *Eurasia Journal of Mathematics, Science and Technology Education* 13, 3 (2016), 893–910.
- [11] Kathryn S McCarthy, Micah Watanabe, Jianmin Dai, and Danielle S McNamara. 2020. Personalized learning in iSTART: Past modifications and future design. *Journal of Research on Technology in Education* 52, 3 (2020), 301–321.
- [12] Nunuy Nurkaeti. 2018. Polya's strategy: an analysis of mathematical problem solving difficulty in 5th grade elementary school. *Edu Humanities| Journal of Basic Education Cibiru Campus* 10, 2 (2018), 140.
- [13] George Polya. 2004. *How to solve it: A new aspect of mathematical method*. Number 246. Princeton university press.
- [14] Alexander Renkl. 1999. Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology of Education* 14, 4 (1999), 477–488.
- [15] Alan H Schoenfeld. 1983. Beyond the purely cognitive: Belief systems, social cognitions, and metacognitions as driving forces in intellectual performance. *Cognitive science* 7, 4 (1983), 329–363.
- [16] NSH Simpol, M Shahrill, HC Li, and RCI Prahmana. 2017. Implementing thinking aloud pair and Polya problem solving strategies in fractions. In *Journal of Physics: Conference Series*, Vol. 943. IOP Publishing, 012013.
- [17] Dirk T. Tempelaar, André Heck, Hans Cuypers, Henk van der Kooij, and Evert van de Vrie. 2013. Formative assessment and learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (Leuven, Belgium) (LAK '13). Association for Computing Machinery, New York, NY, USA, 205–209. <https://doi.org/10.1145/2460296.2460337>
- [18] Yuwalee Thiangthung. 2016. Applying Polya's four-steps and Schoenfeld's behavior categories to enhance students' mathematical problem solving. *Journal of Advances in Humanities and Social Sciences* 2, 5 (2016), 261–268.
- [19] Candace Walkington and Matthew L Bernacki. 2020. Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. , 235–252 pages.
- [20] Huanhuan Wang and James D Lehman. 2021. Using achievement goal-based personalized motivational feedback to enhance online learning. *Educational Technology Research and Development* 69, 2 (2021), 553–581.
- [21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [22] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. <https://doi.org/10.1145/3491102.3517582>
- [23] Erna YAYUK and H Husamah. 2020. The difficulties of prospective elementary school teachers in item problem solving for mathematics: Polya's steps. *Journal for the Education of Gifted Young Scientists* 8, 1 (2020), 361–368.