Can Students Identify the Relevant Information to Solve a Problem?

Lishan Zhang[1,2], Shengquan Yu[1,2*], Baoping Li[1,2] and Jing Wang[1]

[1]Beijing Advanced Innovation Center for Future Education, Beijing Normal University, China // [2]Faculty of Education, Beijing Normal University, China // lishan@bnu.edu.cn // yusq@bnu.edu.cn // libp@bnu.edu.cn // jingwang0422@163.com

[*]Corresponding author

ABSTRACT

Solving non-routine problems is one of the most important skills for the 21st century. Traditional paper–pencil tests cannot assess this type of skill well because of their lack of interactivity and inability to capture procedural data. Tools such as MicroDYN and MicroFIN have proved to be trustworthy in assessing complex problem-solving performances in the dynamic environments representing linear structural equations and finite state automata. In contrast to previous studies, this paper introduces a system that assesses how an individual acquires information to solve real-life problems. Specifically, the system investigated whether fifth-grade students could recognize a situation in which additional information is needed, acquire the relevant information, and finally apply it to solve their problems. By running an experiment with a total of 32 fifth-grade students, we found that the students were usually able to recognize situations when they needed additional information. However, students sometimes spent too much time reading irrelevant materials, which was significantly correlated with worse problem-solving performance ($r = 0.417$, $p = .018$).

Keywords

Problem solving, Problem-based learning, Information identification

Introduction

Regardless of their occupation, people need to handle different types of problems every day. Real-life problems are usually very complex and cannot be solved in a routine manner. Therefore, knowing how to solve these unusual problems has become an essential skill for the 21st century (Greiff et al., 2014a; Griffin, McGaw, & Care, 2012; Neubert, Mainert, Kretzschmar, & Greiff, 2015). Problem solving is not only a skill to deal with real-life situations, but also plays an important role in many learning environments, such as problem-based learning (PBL) (Merritt, Lee, Rillero, & Kinach, 2017).

Problem solving is the process of finding a method to achieve a goal from an initial state. Depending on the domain, the initial state, goal, and means can be very different. Therefore, domain expertise usually plays a dominant role in an individual's problem-solving performance, as described by Chi and Glaser (1983). Well-educated adults may exhibit equally good domain-general problem-solving strategies. However, young students clearly have different levels of competence in solving complex domain-general problems (Findings, 2014), which has fostered a great demand for teaching domain-general problem-solving skills to these students (Greiff et al., 2014a).

In addition, a domain-general problem-solving ability is essential for PBL, which is an effective method for enhancing deep learning (Dolmans, Loyens, Marcq, & Gijbels, 2015). In PBL, students are expected to learn by solving open-ended problems. Open-ended questions are usually complex and poorly structured; therefore, students should have strong self-regulation abilities to be successful in this type of learning. In other words, it is not appropriate to assume that every student is well prepared for PBL. Thus, PBL often involves extensive tutor facilitation, which some educators find difficult and frustrating (Wood, 2003). When facilitation is absent or insufficient, PBL is sometimes found to be less effective than traditional lecture learning (Kirschner, Sweller, & Clark, 2006). Recently, many big cities in China, such as Beijing and Shanghai, have developed a strong trend of adopting PBL in elementary schools. However, as the normal size of a Chinese elementary school class is 40 students, it is not possible for a teacher to help all students efficiently. Therefore, it is even more important to assess students' abilities in conducting PBL-related activities in China to enable teachers to have a better sense of which of their students may need the most help.

Our main objective in problem-solving assessment is to check whether students are ready for PBL; thus, we need an assessment tool fit for this purpose. Some tools have been developed for domain-general problem-solving assessment. Among these tools, MicroDYN and MicroFIN (Schweizer, Wüstenberg, & Greiff, 2013) are the best established. When using these tools, students are required to investigate the complex dependencies of several variables within a dynamically changing situation. Indeed, these tools can be used to reliably assess problem-solving abilities by describing problem situations with linear structural equations and finite state automata.

However, PBL also requires other dimensions of problem-solving ability. Previous studies have shown that students often failed in PBL because of its high cognitive load (Kirschner, Sweller, & Clark, 2006; Sweller, 1988; Tuovinen & Sweller, 1999). One important of cognitive load-related factor is reading literacy. PBL usually involves a great deal of information searching and selection, which requires students to be able to decide what they need to know to resolve their problem (Holliday, 2006). In this context, students need to do much more discontinuous reading than continuous reading. However, Chinese students exhibit worse performance in discontinuous reading than students from other countries according to the report from the 2009 Programme for International Student Assessment (PISA) (Organisation for Economic Co-operation and Development (OECD), 2010). We are concerned that many Chinese elementary school students may experience difficulties in actively searching for the relevant information to solve their problems. Thus, we built an assessment system to evaluate our students' problem-solving ability and their ability to pay attention to different types of information. We will introduce this system in this paper and report how students' attention to information is related to their problem-solving performance. By conducting this experiment, we mainly wanted to answer two research questions:
RQ1. Are students able to source the relevant information and apply them in solving real-life problems?
RQ2. Can the attention given to some specific types of information be used as a predictor of problem-solving performance?

The paper is organized as follows. We review the related studies and then introduce our system and the assessment task. We then describe the experimental design and report the results. Finally, we discuss our findings and conclude with some final remarks.

Related work

Complex problem solving in dynamic environments

While being assessed in a simulated dynamic environment, individuals' core competence for problem solving is shown in their ability to determine the complex dependencies among the observable variables. The most well-known project in this field is probably MicroDYN, which was developed by Schweizer et al. (2013), where students are expected to determine the dependencies of variables in a complex system by manipulating the variables and observing their effects in a dynamic environment. A recent study found that the assessed skill has a strong correlation with traditional reasoning test performances, but was also an independent dimension of ability (Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016; Greiff & Neubert, 2014b). Previous studies have shown the value of this perspective. However, this perspective is not sufficient to explain all types of problem-solving activities. Many problems do not contain complex variable dependencies, but instead need students to distinguish related information from a great number of documents. To solve these problems, problem solvers must be clear about what information they want and purposively search for this information.

Guidance required for problem-based learning

Both problem- and project-based learning can be abbreviated as PBL. However, PBL is used to abbreviate problem-based learning in this paper. PBL has attracted research attention for many years (Holliday, 2006; Kirschner, Sweller, & Clark, 2006; Merritt, Lee, Rillero, & Kinach, 2017) and has been praised because it can motivate students and trigger deep thinking (Merritt, Lee, Rillero, & Kinach, 2017). However, PBL has simultaneously received much criticism (Holliday, 2006; Klahr & Nigam, 2004; Kirschner, Sweller, & Clark, 2006; Patel, Groen, & Norman, 1993; Tuovinen & Sweller, 1999). The main issue for PBL is that students need assistance and facilitation during the learning process and have difficulty in self-regulating their learning. Students, especially those with low levels of prior knowledge, can easily develop a great burden on their cognitive load (Sweller, 1988) and fail to distinguish what they really need to know and learn. This may result in students learning useless or even incorrect knowledge and concepts (Harris & Graham, 1994). Therefore, teachers need to be highly involved in PBL to motivate, regulate, and provide hints to their students so that they do not get lost in looking for the relevant information. Recently, researchers have started to integrate PBL with direct instruction, which seems to be a viable solution (Holliday, 2006; Jalani & Lai, 2015). Direct instruction usually teaches learning strategies for students at the metacognitive level and helps them to avoid making errors in their studies. The direct instruction content is mainly made based on the teacher's experience. By developing this assessment tool, we hope to provide teachers with a better sense of what should be taught to their students.

Typical problem-based learning practice

PBL is most widely used in medical education; therefore, this domain is used as an example to review how PBL may be implemented in practice. Students are usually given a patient's problem related to the skills to be taught. They then learn the related material through self-directed studies to solve the patient's problem (Distlehorst, Dawson, Robbs, & Barrows, 2005). This learning process often involves group study and discussion. The group size can be varied by cases, but usually does not exceed eight students. Students are sometimes grouped by their interests so that they can search for the problem-related information based on their preferences (Distlehorst, Dawson, Robbs, & Barrows, 2005). Teacher facilitation is provided to help students find the relevant information and improve group discussions (McParland, Noble, & Livingston, 2004). In terms of assessment, the studies usually focus on information acquisition, self-regulation, and collaborative study (Distlehorst, Dawson, Robbs, & Barrows, 2005). Teachers usually conduct the assessment by grading students' submitted reports. Dickison et al. (2016) assessed nursing clinical judgement via a computer-simulated environment by analyzing the students' recorded behaviors. In that system, students can gather information by looking at their patient's temperature, medical report, lab results, and vital signs. The students then need to determine the appropriate treatment based on the gathered information. Therefore, the identification of relevant information is one of the most important steps in PBL.

Behavioral analysis

An evidence-centered design (ECD) should be adopted to create a system to support the analysis of students' intentions based on behavioral data. The fundamental design concept was first described by Mislevy (1994). An ECD defines an assessment framework to ensure that evidence is gathered appropriately to be able to interpret the underlying purpose of the assessment. Many common design features are shared by tutoring and assessment systems, although they may be implemented based on their own interpretation of an ECD (Shute, Wang, Greiff, Zhao, & Moore, 2016; Shute, 2011). The very first aspect of adopting this framework is to define the domain modeling, i.e., to clarify the skills to be assessed and sketch the relationships among the proficiencies of the skills, tasks, and evidence. The next main step is to detail the relationships. This process can be factored into three models: student, evidence, and task models. When ECD is applied in an interactive environment-based assessment, a task model defines the story line and how the tasks can elicit students to interact (Halverson & Owen, 2014). An evidence model then describes how the interactions should be analyzed. The evidence model can be based on either statistics or rules. For example, Zhang et al. (2014) inferred students' proficiencies in a meta-strategy using a set of rules and a special sequence of behaviors. Using a hidden Markov model, Schwartz et al. (2009) explored how students interacted with teachable agents. Bayesian networks are widely known statistical analysis models used in many tutoring systems (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). Our system adopts an ECD framework that mainly uses rules to analyze students' problem-solving competences.

The assessment system

To succeed in PBL, we expect that given a real-life situation, students can at first identify what factual knowledge they need to know to solve the corresponding problem, and then search for and solve the problem by using the relevant factual knowledge. We provided problem-irrelevant and -relevant information to the students to evaluate how well they could deal with the related cognitive load. The relevant and irrelevant information were presented together, but different types of information were put into separate documents. Skilled students are expected to keep their cognitive load low by focusing only on the task-related materials. Because the students' patterns of accessing different types of information can reflect how they value the information (Johnson, Häubl, & Keinan, 2007), we inferred the students' focus by observing their recorded behavioral patterns. Therefore, we designed and implemented the system in such a way that the students' information-searching and problem-solving behaviors can be recorded and detected easily. The system includes one assessment task containing four test items. The relevant and irrelevant documents are stored in a virtual library, which is named the "materials center." Students must visit the materials center and read the relevant materials to solve the related problems. We are interested in how students value different information in the materials center. The most straightforward method is to directly ask students to select the information that they think is important. However, this design will interrupt the students' natural problem-solving processes. Thus, we non-intrusively infer how they value the information by analyzing the log files. The rest of this section discusses the user interface, details the contents of the assessment task, and describes the types of behaviors that we recorded.

The user interface

To support the structure of the test items, the system provides some common functionalities and utilities. To allow students to access the relevant materials conveniently, a navigation bar is placed at the right-hand side of the area where the test items are displayed (Figure 1). All associated materials for the current assessment task are stored in the component, "materials center," which is located in the navigation bar. The materials center contains not only the relevant materials but also some irrelevant ones. Therefore, the follow-up data analysis could reveal how students chose different materials. As soon as a student clicks on the materials center, a new window pops up that shows a list of the materials (Figure 2, left). Each entry in the list shows both the title and a brief description of the document. When a student clicks on the name of the document, its details are displayed (Figure 2, right). The student can click on the back button to return to the list panel and access other materials. Students can also close the materials center whenever they want by clicking on the close button. High-achieving students are supposed to be very clear about what they are looking for and can thus locate the relevant document quickly. By comparison, low-achieving students could flounder and may randomly access many irrelevant documents. As soon as the students finish a test item, they can proceed to the next one.

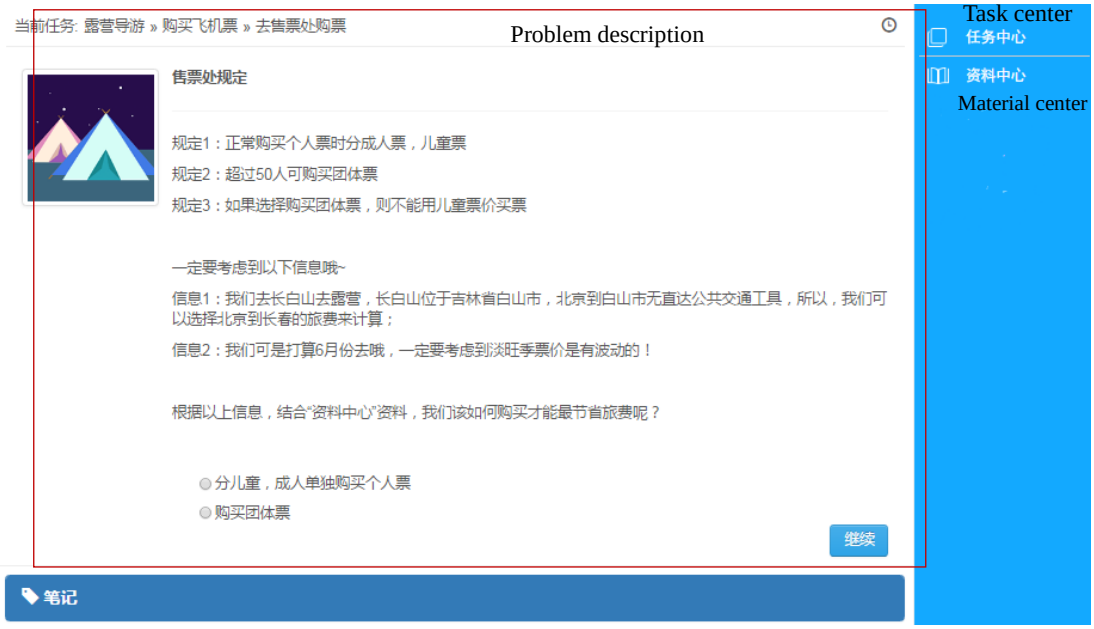

Figure 1. A sample test item (test item 1)

Figure 2. Materials center: List of materials (left) and contents of a single item (right)

The assessment task

To help the students learn the user interface, they must perform an introductory task before performing the actual assessment task. The introductory task is very simple: students need to calculate how much time would be taken to fly from Beijing to Washington, DC. The airplane speed has been given in the description of the problem. The distance between the two cities is included as a document in the materials center, which also includes other irrelevant material. The introductory task asked students to first locate the relevant material, and then perform

the calculation. The students must finish both parts of the introductory task to complete it. Otherwise, the system prompts the students to recognize which part they did not finish.

Table 1. Summary of test items

Item name/Feature	Item type	Has relevant materials?
Ticket purchase	Multiple choice	Yes
Tent capacity calculation	Interactive	Yes
Tent assignment	Interactive	No
Food and water supply	Fill-in-blank	Yes

The actual assessment task contains four test items and uses camping as the story line. The four standalone test items share the same background story line. The features of these four test items are summarized in Table 1. In addition to the traditional types of items, such as multiple choice and fill-in-blank, students need to interact with a simulated environment to solve problems in two test items. These items are "interactive." For each test item that requires students to search for relevant information, we provided three types of materials at three different levels of relevance: strong, medium, and weak. Students are required to obtain materials with strong relevance to solve the problem. Materials with medium relevance are related to the problem situation and materials with weak relevance are completely irrelevant to the problem. Some material examples are listed in Table 2. The remainder of this section describes each test item in detail.

Table 2. Test items and their materials with strong, medium, or weak relevance

Test item	Level of relevance	Brief description of example materials
Ticket purchase	Strong	Flight ticket from Beijing to Liaoning
	Medium	Required camping equipment
	Weak	Things to be aware of while camping at sand beach
Tent capacity	Strong	Comfortable size for an individual while living in a tent
	Medium	How to fix a broken tent
	Weak	The average size of living space for an individual in Beijing
Food and water supply	Strong	The average amount of food that needs to be consumed in a camping day
	Medium	How to select a camping spot
	Weak	How to select an appropriate tent

Test item 1: "Ticket purchase"

In the first test item, students are expected to read the relevant materials to determine the best way to purchase tickets. Students are told that they plan to go camping outside. They must fly to the camping area because it is far from where they are living. Students are required to decide whether they should purchase group or individual tickets. The problem clearly describes the flight destination. Students should recognize that they need to distinguish the difference between the two purchasing methods, visit the materials center to check the ticket-purchasing rules, and then finally make their decision.

Test item 2: "Calculating tent capacity"

The second test item requires students to calculate the capacity of their tents (Figure 3, left). The students are expected to measure the sides of three tents correctly and then use the methods described in the relevant materials to calculate the capacities of the tents, which differ only in their sizes. To measure a side of a tent, the students need to click on the side, read the length from a pop-up image (Figure 3, right), fill out the value, and close the pop-up image. Each tent has five sides to be measured, but only two of them are useful for the calculation of the capacity. Therefore, the students can potentially make some invalid interactions—i.e., making unnecessary measurements—in this task. Students who read and understood the relevant materials well should be able to avoid performing these invalid interactions. Although making invalid interactions does not harm the students' final performance score, it is a sign of shallow thinking.

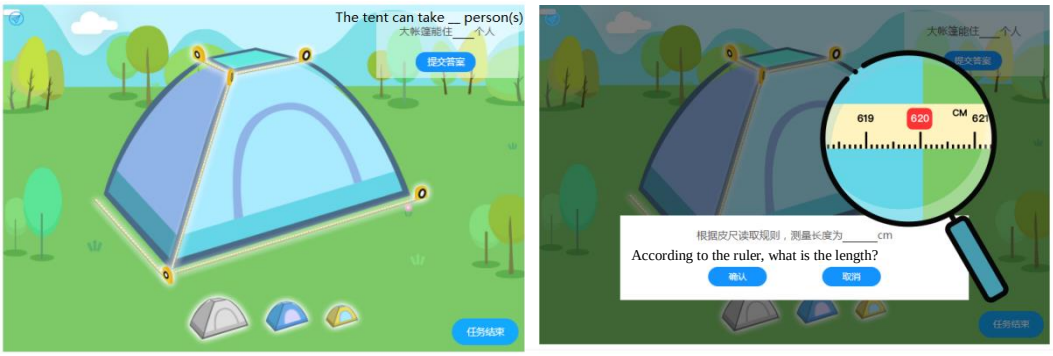

Figure 3. Screenshot of test item 2: "Tent capacity calculation" (left), pop-up image showing the tent side measurement (right)

Test item 3: "Tent assignment"

The third test item does not require students to read any relevant materials. However, the students can visit the materials center if they want. To solve this problem, students need to assign 54 persons to 7 tents (Figure 4). The 54 persons can be categorized into four types, according to their age and gender: i.e., boys, girls, and male and female adults. There are three different types of tents. Each type of tent can contain a limited number of persons. The tent assignment should satisfy three constraints: (1) People in the same tent must be the same gender; (2) There must be at least one adult in a tent; (3) The number of persons in a tent cannot exceed its limit. This problem checks whether students can recognize a situation where they do not need to refer to additional material.

Test item 4: "Food and water supply"

In the last task, students are told that they can only resupply their food and water at the end of each day during their camping trip; therefore, they need to calculate how much food and water they should prepare for each day. The materials center provides the corresponding documents that describe how much food and water an adult usually needs to consume daily. The students must obtain this information to solve the problem. In contrast to the first test item, the students need to type their answer. Because the answers are specific numbers, they can be easily graded automatically.

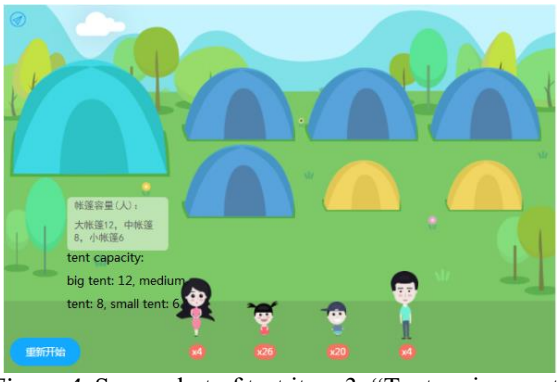

Figure 4. Screenshot of test item 3: "Tent assignment"

Behavioral record

Log files have been widely used to understand students' performance (Greiff, Wüstenberg, & Avvisati, 2015; Kuo & Wu, 2013). Our system records every single interaction, such as clicking on an alternative and accessing the relevant materials. Thus, the behavioral data can be analyzed offline. When a test item is a multiple-choice question, the set of possible interactions is limited and only contains "choosing an alternative" and "accessing different types of materials." However, the number of possible operations a student can make in an interactive test item is potentially much greater and completely depends on the specific problem. To standardize all interactions, except for materials center-accessing behaviors, they are labeled with four different types: correct, incorrect, valid, and invalid. Correct and incorrect behaviors describe the correctness of an interaction, but the

correctness criteria are changed by specific problems. Test item 2 intentionally embedded some unnecessary interactions towards the problem goal. Students can either skip or finish these steps. In test item 2, a necessary interaction is labeled as "valid" and other interactions are labeled as "invalid." For each material-accessing behavior, we recorded the timestamp and the material's corresponding level of relevance. The details of our analysis are described in the next section.

Experiment design and analysis model

Experiment design

All students who participated in our experiment were in the fifth grade. We performed a pilot study with approximately 20 students in the same grade to evaluate the usability of the system and the assessment task. The introductory task was not included initially. As a result, the students in the pilot study had no idea how to use the materials center to help them solve problems. Thus, we created an introductory task to guide the students in accessing the materials center. This introductory task essentially helped students to become familiar with the user interface. The pilot study also helped us identify several software bugs and ensure that the problem description could easily be understood by fifth-grade students. Students were required to finish the introductory task before starting the actual assessment task. They were asked to finish the assessment individually in a class.

Analysis model

The first question we needed to answer was how well students performed when they had to find out part of the facts related to the problems by themselves. Therefore, we calculated the descriptive results of their performance scores on each individual test item. In addition, we wanted to understand how students solve problems by observing how they used the materials center and identify their issues during the problem-solving process. Therefore, we mined the relationship between students' behavioral patterns and performance outcomes on test items. Pearson's correlation coefficient was calculated to measure the relationship. Linear regression was further applied to evaluate whether the students' behavioral patterns could be used to predict their final problem-solving performance.

In analyzing the students' behavioral patterns, we focused on their interactions with the documents in the materials center. Because we categorized all documents into three types for each test item—strong, medium, and weak—we could examine how students allocate their attention to the three types of materials. As soon as students opened a document, the pop-up window would prevent them from performing all the other behaviors (e.g., answering the question or opening another document). We used the time that the student stayed at an opened document to reflect how important the student thought the material was. We assumed that if the students spent at least 5 seconds in reading a document, they probably considered the material as somewhat important. Each reading behavior was labeled as either short-reading or non-short-reading behaviors. Five seconds was used as the threshold to distinguish between these behaviors. Therefore, we classified the reading behaviors into 2 (short-reading and non-short-reading behaviors) × 3 (strong, medium, and weak relevance) = 6 different types.

In addition to the test item scores, the test item "Tent capacity calculation" could also report other information to reflect the students' problem-solving performance. As mentioned earlier, the students can potentially perform invalid work in solving their problems. Specifically, an example of invalid work is measuring tent sides unnecessarily. Because the related information in the materials center clearly states which two sides should be measured to calculate the capacity of tents, the students who acquire the information accurately and are thinking deeply should be able to avoid invalid work (Holliday, 2006). Thus, the percentage of unnecessary sides measured was calculated to reflect whether the students were thinking shallowly.

Results

Descriptive results

As mentioned above, there were four test items in total. Each item had a weight of 1 point. Therefore, the maximum possible score on the test was 4. Thirty-two fifth-grade students participated in the experiment. The average score on the entire assessment task was 1.572 (SD = 0.762). Table 3 reports students' performance and their reading behaviors on each individual test item. Because the test item, "Tent assignment," does not have any

relevant materials, we aggregated the students' performance and their reading behaviors on the other three test items (Table 3). The results indicated that students had low scores while solving the problems that required reading relevant materials, and scored relatively high while solving the problems without relevant materials.

Table 3. Descriptive results of the test items

Task	Ticket purchase	Tent capacity calculation	Tent assignment	Food and water supply	Overall performance without test item 3
Score	0.5 (*SD* = 0.508)	0.25 (*SD* = 0.237)	0.68 (*SD* = 0.413)	0.136 (*SD* = 0.195)	0.89 (*SD* = 0.572)
Number of short-reading strong relevant materials	0.063 (*SD* = 0.242)	0.031 (*SD* = 0. 174)	N/A	0.094 (*SD* = 0. 291)	0.19 (*SD* = 0.464)
Number of short-reading medium relevant materials	0.094 (*SD* = 0. 291)	0.16 (*SD* = 0.44)	N/A	0.063 (*SD* = 0. 242)	0.31 (*SD* = 0.527)
Number of short-reading weak relevant materials	0.125 (*SD* = 0. 415)	0.094 (*SD* = 0.291)	0 (SD = 0)	0.59 (*SD* = 2.16)	0.81 (*SD* = 2.17)
Number of non-short-reading strong relevant materials	0.44 (*SD* = 0.556)	0.75 (*SD* = 0.612)	N/A	0.50 (*SD* = 0.66)	1.69 (*SD* = 0.982)
Number of non-short-reading medium relevant materials	0.56 (*SD* = 0.788)	0.28 (*SD* = 0.514)	N/A	0.094 (*SD* = 0.291)	0.94 (*SD* = 0.899)
Number of non-short-reading weak relevant materials	0.125 (*SD* = 0.545)	0.094 (*SD* = 0.291)	0.15 (*SD* = 0.330)	0.34 (*SD* = 0.988)	0.56 (*SD* = 1.09)

Correlation results

Pearson's correlation coefficients were calculated to determine whether material-reading behaviors were related to student performance. Specifically, for each student, we first aggregated his/her total number of short-reading and non-short reading behaviors on materials with strong, medium, or weak relevance. We then calculated the correlation between the six factors and the students' overall performance except for the task, "Tent assignment." The results are shown in Table 4. The amount of non-short-reading behaviors for materials with weak relevance was the only factor that significantly correlated with problem-solving performance. All the factors were also used to construct linear regression to predict overall performance without test item 3 using a stepwise algorithm. We found that the overall performance can be predicted by the number of non-short-reading weak relevant materials (β = -0.417, R^2 = 0.174) and a constant.

The percentage of unnecessary sides measured in the task, "Tent capacity," was used to reflect students' shallow thinking. The correlation between material-reading behaviors and evidence of shallow thinking was calculated. Among all of the reading behaviors, the number of short-reading behaviors for materials with weak relevance was the only one that significantly correlated (r = 0.398, p = .024) with evidence of shallow thinking. This means that students who quickly scanned a lot of irrelevant materials also tend to simply do work without thinking about the reason.

Table 4. How reading behaviors correlate with overall task performance and the sign of shallow thinking

Reading behaviors	Short-reading strong relevant materials	Short-reading medium relevant materials	Short-reading weak relevant materials	Non-short-reading strong relevant materials	Non-short-reading medium relevant materials	Non-short-reading weak relevant materials
Correlation with task performance	-0.183 $p = .317$	-0.198 $p = .279$	-0.298 $p = .091$	0.117 $p = .523$	0.097 $p = .597$	-0.417* $p = .018$
Correlation with the sign of shallow thinking	0.156 $p = .367$	0.153 $p = .404$	0.398* $p = .024$	0.102 $p = .580$	0.015 $p = .934$	0.015 $p = .937$

Note. $^{*}p < .05$.

Discussion

The results showed that the students did not perform very well in general. Although this type of assessment task is very different from those that they are usually assigned, these problems themselves are not very difficult. Teachers believe that fifth-grade students have enough prior knowledge to solve them. Probably because the problems were somewhat ill-structured—i.e., part of the related factual knowledge was stored in the materials center—and students sometimes failed to locate the information precisely. Most students could not answer the questions correctly. This claim is supported by our data analysis. The more students spent their time in reading completely irrelevant materials, the more likely these students would receive a bad score on their overall performance. This result is consistent with the theory of cognitive load (Sweller, 1988). Students should efficiently use their limited working memory to handle challenging problems. Reading and processing irrelevant information might occupy too much of students' working memory during problem solving, and increase their cognitive loads. As PBL will make students explore an even bigger problem space than the one in the experiment, teachers should provide strong facilitation to guide their students throughout PBL. Note that this elementary school is one of the best schools in Beijing. Therefore, we might expect that more facilitation is required in other schools. However, the results showed that most students could recognize the situation where they did not need to search for any additional facts. Thus, students were not completely lost. We believe that it is practical to train students to recognize and focus on the relevant information when necessary. As we have already implemented a system that is able to detect different types of reading behaviors, one of the next steps is adding intervention to persuade students to keep their focus on situation-related information to reduce their cognitive load. Hopefully, this kind of training can help students solve these types of ill-structured problems more efficiently and successfully complete their PBL activities.

Another explanation for the students' poor performance is that they do not perform well in discontinuous reading, as shown in the results reported by the 2009 PISA (OECD, 2010). Although the assessment results were from several years ago, fundamental changes have not been made in the Chinese reading instruction system. Most of the reading practices, especially those in elementary schools, are still well structured so that students can easily digest the learning contents. Although the test items in this assessment task clearly state what needs to be done, the given contents are distributed in different locations and mixed with other irrelevant information, which makes the assessment task require a more discontinuous reading ability instead of continuous reading ability.

The results suggest that many students are not good at acquiring factual knowledge by themselves to solve real-life problems. Therefore, teachers should be very careful in conducting PBL in elementary schools. Teachers are expected to provide a great amount of facilitation to their students. This paper does not try to prevent other teachers from using PBL in elementary schools, but instead remind them of the related issues. Notably, previous studies have shown that after students become comfortable with PBL, they no longer need too much facilitation (Holliday, 2006; Jalani & Lai, 2015).

In addition to students' problem-solving outcomes, we also looked at how much invalid work they performed. Students who paid too much attention to irrelevant information tend to perform more invalid work. These students simply tried to work out every piece of information without thinking carefully. They might try their best to solve the problems but just not solve the problems using the correct method. It is possible that these students

may study very hard in PBL, but still not learn the appropriate lessons from the task. Making students realize and focus on what they need to know seems straightforward, but is not easy to implement in practice. Zhang et al. (2014) found that high school students also tend to simply work out every piece of work whether or not it is useful for solving their problem.

After identifying these issues, how can we best help the students? First of all, our assessment system can be modified to train students to self-regulate their problem-solving process and reduce their cognitive loads by giving them hints to avoid placing too much of their attention on irrelevant materials. Secondly, schools should urge their teachers to encourage their students to practice more discontinuous reading before doing PBL to train their ability to synthesize information.

In PBL, students often need to seriously think about what they need to know. While making these decisions, students have to relate the current situation and information to their prior knowledge, which is considered as a deep approach to learning (Crooks & Alibali, 2013; Dolmans et al., 2015). The design of our system makes it possible to detect the moment a student decides what is necessary to know. If PBL's learning materials can be arranged according to our system, i.e., put all the learning materials into the materials center and label each document, the system may also help students during the PBL process.

Conclusion and limitations

This paper introduced a technology-enhanced problem-solving ability assessment system that can track how students solve problems by analyzing their reading behaviors. In our experiment, most of the fifth-grade students were able to recognize whether to search relevant information for solving a specific task, but often failed in locating the relevant information when needed. The results also suggested that the distractions of irrelevant information could lead to bad problem-solving performance as well as shallow thinking.

The study clearly has several limitations. Firstly, the sample size is relatively small, and the participants are from the same class. Readers should be careful while generalizing our conclusion into a bigger population. Secondly, this study focused on exploring how well students identified text information to solve problems. Students might perform differently when the information is presented as figures, tables, or other formats. Thirdly, the study cannot externalize how students process the identified information in their minds. This is an aspect of our future research.

Acknowledgments

This work was supported in part by China Postdoctoral Science Foundation under projects number 2017M610054, and by Philosophy and Social Sciences Research of Chinese Ministry of Education under projects number 16JZD043.

References

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

Chi, M. T., & Glaser, R. (1983). *Problem solving abilities* (Technical report No. 8). Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.

Crooks, N. M., & Alibali, M. W. (2013). Noticing relevant problem features: Activating prior knowledge affects problem solving by guiding encoding. *Frontiers in Psychology, 4*(1), Article 884. doi:10.3389/fpsyg.2013.00884

Dickison, P., Luo, X., Kim, D., Woo, A., Muntean, W., & Bergstrom, B. (2016). Assessing higher-order cognitive constructs by using an information-processing framework. *Journal of Applied Testing Technology, 17*, 1-19.

Distlehorst, L. H., Dawson, E., Robbs, R. S., & Barrows, H. S. (2005). Problem-based learning outcomes: The Glass half-full. *Academic Medicine, 80*(3), 294-299.

Dolmans, D. H., Loyens, S. M. M., Marcq, H., & Gijbels, D. (2015). Deep and surface learning in problem-based learning: A Review of the literature. *Advances in Health Sciences Education, 21*(5), 1087-1112.

Findings, K. (2014). *PISA 2012 results: Creative problem solving: students' skills in tackling real-life problems* (Vol. V). Paris, France: Organisation for Economic Co-operation and Development (OECD).

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., & Graesser, A. C., & Martin, R. (2014a). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13*(3), 74-83.

Greiff, S., & Neubert, J. C. (2014b). On the relation of complex problem solving, personality, fluid intelligence, and academic achievement. *Learning and Individual Differences, 36*, 37-48.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A Showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92-105.

Griffin, P., McGaw, B., & Care, E. (Eds.) (2012). *Assessment and teaching of 21st century skills.* Dordrecht, The Netherlands: Springer.

Halverson, R., & Owen, V. E. (2014). Game-based assessment: An Integrated model for capturing evidence of learning in play. *International Journal of Learning Technology, 9*(2), 111-138.

Harris, K. R., & Graham, S. (1994). Constructivism: Principles, paradigms, and integration. *Journal of Special Education, 28*(3), 233-247.

Holliday, W. G. (2006). A Balanced approach to science inquiry teaching. In L. B. Flick & N. G. Lederman (Eds.), *Science & Technology Education Library* (pp. 201-217). Dordrecht, The Netherlands: Springer.

Jalani, N. H., & Lai, C. S. (2015). The Example-problem-based learning model: Applying cognitive load theory. *Procedia - Social and Behavioral Sciences, 195*, 872-880.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A Query theory of value construction. *Journal of Experimental Psychology Learning Memory & Cognition, 33*(3), 461-474.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An Analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 4*1(2), 75-86.

Klahr, D., & Nigam, M. (2004). The Equivalence of learning paths in early science instruction: Effect of direct instruction and discovery learning. *Psychological science, 15*(10), 661-667.

Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A Comprehensive view on different facets of intelligence and school grades. *Intelligence, 54*, 55-69.

Kuo, C. Y., & Wu, H. K. (2013). Toward an integrated model for designing assessment systems: An Analysis of the current status of computer-based assessments in science. *Computers & Education, 68*(1), 388-403.

McParland, M., Noble, L. M., & Livingston, G. (2004). The Effectiveness of problem-based learning compared to traditional teaching in undergraduate psychiatry. *Medical Education, 38*(8), 859-867.

Merritt, J., Lee, M. Y., Rillero, P., & Kinach, B. M. (2017). Problem-based learning in K–8 mathematics and science education: A Literature review. *Interdisciplinary Journal of Problem-Based Learning, 11*(2). doi:10.7771/1541-5015.1674

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 58*, 79-85.

Neubert, J. C., Mainert, J., Kretzschmar, A., & Greiff, S. (2015). The Assessment of 21st century skills in industrial and organizational psychology: Complex and collaborative problem solving. *Industrial and Organizational Psychology, 8*(2), 238-268.

Organisation for Economic Co-operation and Development (OECD). (2010). *PISA 2009 results: Executive summary.* Paris, France: OECD Publishing.

Patel, V. L., Groen, G. J., & Norman, G. R. (1993). Reasoning and instruction in medical curricula. *Cognition & Instruction, 10*(4), 335-378.

Schwartz, D. L., Chase, C., Chin, D. B., Oppezzo, M., Kwong, H., Okita, S., Biswas, G., Roscoe, R., Jeong, H., & Wagster, J. (2009). Interactive metacognition: Monitoring and regulating a teachable agent. In *Handbook of metacognition in education* (pp. 340-359). New York: NY: Routledge.

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning & Individual Differences, 24*(2), 42-52.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction, 55*(2), 503-524.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106-117.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285.

Tuovinen, J. E., & Sweller, J. (1999). A Comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91*(2), 334-341.

Wood, D. (2003). The Why? what? when? and how? of tutoring: The Development of helping and tutoring skills in children. *Literacy Teaching & Learning, 7*(1-2), 1-30.

Zhang, L., VanLehn, K., Girard, S., Burleson, W., Chavez-Echeagaray, M. E., Gonzalez-Sanchez, J., & Hidalgo-Pontet, Y. (2014). Evaluation of a meta-tutor for constructing models of dynamic systems. *Computers & Education, 75*, 196-217.