Computers & Education 95 (2016) 1–18

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique

Yao-Ting Sung [a], Chia-Ning Liao [a, *], Tao-Hsing Chang [b], Chia-Lin Chen [c], Kuo-En Chang [c]

[a] Department of Educational Psychology and Counseling, National Taiwan Normal University, 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan
[b] Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, No.415, Jiangong Rd., Sanmin Dist., Kaohsiung City 807, Taiwan
[c] Department of Graduate Institute of Information and Computer Education, National Taiwan Normal University, 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan

ARTICLE INFO

Article history:
Received 10 November 2014
Received in revised form 12 December 2015
Accepted 15 December 2015
Available online 18 December 2015

Keywords:
Elementary education
Evaluation methodologies
Intelligent tutoring systems
Teaching/learning strategies

ABSTRACT

Studies on teaching of reading strategies have found that summarizing is of tremendous help to reading comprehension. However grading students' summary writings is laborious, but given the importance of summarizing, an effective summarizing learning module is important. This study developed an automatic summary assessment and feedback system based on Latent Semantic Analysis (LSA) to provide score, concept and semantic feedback, and then investigated the effects of concept and semantic feedback on the writing of summaries by students in the sixth grade. The design involved two between-subject factors: semantic feedback (with, without) and concept feedback (with, without). 120 sixth-grade students from an elementary school were recruited for the study, and then were randomly assigned to each group. The overall results demonstrated the effectiveness of the proposed system in improving the summary writing skills of students. The effects of semantic feedback and concept feedback were also discussed.

1. Introduction

Using learning strategies to enhance students' literacy has been recognized as an important approach of teaching (Block & Pressley, 2002; Lan, 2015; Sung, Wu, Chen, & Chang, 2015), among the proposed strategies, summarization is one of the most recommended strategies for developing reading and writing skills (Chang, Sung, & Chen, 2002; Lenhard, Baier, Endlich, Schneider, & Hoffmann, 2013).

A summary is meant to express the important ideas in a text as succinctly as possible. Thus, summary writing tends to be more constrained than open-ended writing styles. The author must understand the text, identify and compile the important points, and finally restate the content in a more concise form in their own words (Kintsch, 1990; Kintsch & Van Dijk, 1978). A summary is the product of summary writing, and has been considered as the representation of a reader's understanding or

* Corresponding author. 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan
E-mail address: artning0905@gmail.com (C.-N. Liao).

http://dx.doi.org/10.1016/j.compedu.2015.12.003

knowledge about a text and has therefore been used as a measure of reading comprehension (Head, Readence, & Buss, 1989). In addition, researchers seeking to develop effective reading strategies (Bean, Singer, Sorter, & Frazee, 1986; Bean & Steenwyk, 1984; Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Malone & Mastropieri, 1992; Weisberg & Balajthy, 1990) have found that instruction or training in summary writing can be of tremendous help to learning performance outcomes, such as retention and comprehension.

Summary writing can be viewed as a helpful learning activity; however, grading students in their summary writing abilities and providing individual feedback is very difficult and time-consuming for teachers. Consequently, many teachers opt not to assign many assignments based on summarization in order to save time for other important tasks. As a result, most students are not provided sufficient opportunities to engage in this effective learning activity. Computer-assisted assessments could be used to assist teachers in the grading of summaries while automatically providing students tailor-made feedback immediately upon completion of their assignments. In this paper, we propose an ensemble approach for improving students' summary writing ability. This approach integrates traditional and newly developed methods: Two traditional summary assessment indices, the "Efficiency of Summarization" by Garner (1982), and "the proportion of important idea units (IMUPIU/IU)" by Head et al. (1989) were used along with two techniques developed recently, namely latent semantic analysis (LSA; Landauer, Foltz & Laham, 1998), an effective automatic summary evaluation method, and concept mapping (Brunt & Karpicke, 2014; Chang et al., 2002), a kind of spatial learning strategy.

The following literature review includes an overview of the proposed summary assessment methods, automatic summary assessment techniques based on LSA, and spatial learning strategies.

1.1. Summary assessment

Numerous methods have been devised for the evaluation of learner performance in regards to summarization, such as Garner's (1982) Efficiency of Summarization, Head et al. (1989) IMUPIU/IU, and LSA-based automatic assessment (Landauer et al., 1998).

Garner (1982) presented the Efficiency of Summarization, which refers the "proportion of number of judged-important ideas included to total number of words in each summary" (p. 275). In that study, 16 graduate students were invited to rate the importance of every sentence in the target text about Dutch elm disease. Then, three important ideas (causes, signs, and remedies) were taken from the target text according to the rated scores of sentences. Finally, Garner and her assistant then identified the three most important ideas in each summary. The number of important ideas and words used in each summary were then counted for use in calculating the summarization efficiency of the students.

Head et al. (1989) used the proportion of important idea units (IMUPIU/IU) as a summary score, which was tallied according to the ratio of the number of important idea units divided by the total number of idea units in the summary. The target passage in that study was parsed into 63 idea units, and assigned one of four scores according to importance, from 1 (least important) to 4 (most important). Summaries were compared with a template of idea units from the target passage for the calculation of IMPIU/IU.

Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) first proposed a statistical method called LSA for indexing documents and retrieval information. The rationale behind LSA is the use of a 2-D term-to-document occurrence matrix for the presentation of relationships between words (terms) and texts (documents). Each element in the matrix refers to the frequency of a term appearing in a document. The matrix is then transformed using singular value decomposition (SVD) in order to construct a semantic structure with fewer dimensions. The semantic space preserves only essential semantic relationships between words and texts rather than surface syntactic features; therefore, Deerwester et al. (1990) named this approach latent semantic analysis.

Landauer and the LSA research group introduced the LSA method to the fields of language research (Landauer et al., 1998) and automatic summary assessment (Foltz, Kintsch, & Landauer, 1998). They built a 60,768 $\times$ 30,473 term-to-document occurrence matrix using the Grolier Encyclopedia as a textual input, and then transformed the occurrence matrix using SVD and dimension reduction methods into a semantic space with approximately three hundred dimensions. The semantic space made it possible to conduct analysis on the meaning of every word, sentence, passage, or text, which could be projected into the semantic space to obtain a vector. The proximity (cosine value) between two vectors is used as an indicator of the similarity between two words, sentences, or texts. The higher the cosine value, the greater the similarity. Because LSA can be used in this way, it can also be used to measure the similarity between source texts and summaries.

1.2. Summary assessment and feedback techniques

Automatic feedback and assessment systems might lessen the work load of teachers when they teach summary writing. Kintsch et al. (2000) employed LSA in the development of two automatic feedback systems for the writing of summaries. These systems were referred to as "State the Essence" and "Summary Street." In both, LSA is used to calculate the relationships between student summaries and source texts, whereupon feedback is immediately relayed to students. The general feedback information provided by these systems deal with (a) spelling, (b) length, (c) an overall score, and (d) the adequacy of section coverage and overall content coverage. In addition, the students could also request checks for (e) redundancy, and (f) relevance.

With State the Essence, they found that the LSA grading system did not differ significantly from how real teachers would grade summaries. The second system, Summary Street, was a slight upgrade to State the Essence in that it provided a few additional methods of feedback, such as adding a graphic user interface. They then compared Summary Street to traditional teaching methods, and found that Summary Street produced superior learning achievement, particularly when dealing with difficult subject matter (Kintsch et al., 2000).

Wade-Stein and Kintsch (2004) performed an additional study to test how feedback affected the effects of Summary Street. They found that feedback caused students to spend more time working on their summaries, and that the end product was superior to groups with minimal feedback (just spelling and length.) Further, they found that this feedback demonstrated a lasting effect on summary writing ability because student ability did not decline when the feedback system was later minimized.

Landauer, Lochbaum, and Dooley (2009) integrated Summary Street and the Intelligent Essay Assessor (IEA; Landauer, Laham, & Foltz, 2003) into WriteToLearn, which employs a Summary Street subsystem capable of providing feedback. This feedback encompasses the following:

> (a) content: how well the summary covers the gist and key points of each section of the reading, (b) length: whether the summary has been adequately condensed from the original text, (c) copying: whether students have copied too much from the original text, (d) spelling: which words may be misspelled, (e) redundancy: whether there are repetitive sentences that could be combined, and (f) irrelevancy: whether there are unrelated sentences that could be omitted. (Landauer et al., 2009, p. 47)

The IEA subsystem provides automated writing assessment capable of automatically evaluating the overall quality of summaries as well as various specified features, including the effective use of sentences, grammar usage, word choice, semantic coherence, and so on. At the same time, LSA technology enables IEA to simulate the process of a human reader in comparing two articles in order to determine the completeness of a student summary. The WriteToLearn system enables teachers to assign assignments according to the skills of their students and monitor their activities and progress. Students can also make use of the immediate feedback to revise their summaries. In interviews conducted by Landauer et al. (2009) students with experience using the system indicated that the feedback function enhanced their summary writing and teachers reported that the system reduced the burden they faced in having to read and mark summaries.

Summary assessment and feedback technologies such as Summary Street and WriteToLearn promote students' learning performance by providing the students with brief feedback. This type of instruction or learning method differs from regular instruction in that it enables learners not merely to passively accept teachers' knowledge or materials but rather to undertake self-regulated learning.

Self-regulated learning means students can learn by actively managing their own cognition, motivation, and behavior. During this process, learners enhance performance and achieve their study objectives (Zimmerman, 1998; 2001; 2002b). Zimmerman (2002b) proposed a cyclical model of self-regulated learning, which can be divided into three phases: forethought, performance, and self-reflection. In the forethought phase, learners establish specific study objectives and suitable learning strategies for themselves based on their self-efficacy, intrinsic interest and goals of study. In the performance phase, learners attempt to execute study strategies, focus their attention on tasks, improve their awareness of their own behavior, and monitor the result of implementation. In the self-reflection phase, based on outside feedback and evaluation obtained after completion of study tasks, learners engage in self-reflection and readjust study objectives or learning strategies. Subsequently, learners' reflection is fed back into the next cycle for forethought, and in this way, the self-regulated learning process continues to cycle, promoting improved learning performance.

Studies have stated that self-regulated learning has a positive influence on learning performance (Azevedo, Cromley, Winters, Moos, & Greene, 2005), a finding which has been confirmed in various fields related to the learning process. Being aware of the status of one's own progression, such as obtaining an evaluation for one's own music, paintings, or essays, can guide one on how to improve and is a key factor influencing the decision regarding whether to persist in the original strategy or not. However, in actual practice, very few teachers can apply self-regulated learning to guide students to become independent learners (Zimmerman, 2002b).

During learners' summary writing process, an automatic summary assessment and feedback system can continually assess learners' summaries and provide feedback. This system can help learners self-evaluate and monitor their study performance, helping them set objectives and strategies more appropriate for future study. In this way, students can improve their learning performance through continuous self-regulated learning.

However, the processes of self-monitoring and adjusting of learning strategies based on feedback consume a substantial amount of time and energy (Zimmerman, 2002a). The question of whether the feedback information provided by the Summary Street and WriteToLearn systems can bring satisfactory benefits remains to be addressed quantitatively.

Additionally, through a review of the literature concerning the development of Summary Street, we can see that sometimes a large amount of feedback information does not necessarily lead to a satisfactory improvement in learning performance. Kintsch et al. (2000), in the context of a practice where they attempted to provide further feedback on redundancy and relevance in addition to general information regarding spell check, length, an overall score, and the adequacy of section coverage and overall content coverage, found that participating students were overwhelmed by excessive information loads. Therefore, they simplified the feedback items for State the Essence. Kintsch et al. (2000) only provided a few measures: a point score (0–100 points), length (too short, too long, or about right), an evaluation of the content of each section (good, ok,

needs improving, or missing), and listing the weakest section along with a hyperlink to that section of the source text in one system version of State the Essence. However, they added feedback on redundancy and relevance back in to the subsequent Summary Street and WriteToLearn, but have yet to justify this with experimental data.

The studies of Wade-Stein and Kintsch (2004) and Landauer et al. (2009) supported the idea that, compared with traditional instruction methods, automatic summary assessment and feedback systems such as Summary Street and WriteToLearn can significantly promote students' summary writing performance, while reducing the burden on teachers. Both the Summary Street and WriteToLearn systems provide students massive feedback regarding various types of information, such as spell check, length, an overall score, the adequacy of section coverage and overall content coverage, along with more specific information, on matters such as redundancy and relevance, etc. Nevertheless, previous studies (Kintsch et al., 2000; Landauer et al., 2009; Wade-Stein & Kintsch, 2004) have never clarified the question of whether more feedback concerning redundancy and relevance has a greater promoting effect on students' summary writing performance or, in contrast, a drag-down effect on student performance because of the added cognitive burden.

Thus, the first objective in this study was attempting to address the following question: Will more detailed feedback regarding relevance provided in addition to general feedback have a significant promoting effect on students' summary writing performance?

1.3. Spatial learning strategies

Summaries represent the gist of a text in linear terms. Spatial learning strategies can also be used to organize and represent the knowledge or conceptual structure of a text within the minds of learners; however, spatial strategies describe this knowledge in spatial rather than linear terms.

There are at least three kinds of spatial strategies: graphic organizers, knowledge maps, and concept maps. They all represent concepts and the relationships between concepts in a given knowledge domain with a spatial configuration of nodes and links. The hierarchical structure of the knowledge makes it easier to retain and retrieve.

Considerable research has demonstrated the benefits of spatial learning strategies on learning performance (Chang et al., 2002; Griffin, Malone, & Kameenui, 1995; McCagg & Dansereau, 1991; Stull & Mayer, 2007). For example, McCagg and Dansereau (1991) found that when undergraduate students used knowledge maps while studying statistics and physiological psychology they learned physiological psychology better than when they did not.

Additionally, Chang et al. (2002) designed three concept-mapping approaches (i.e. map-generation, map-correction, and scaffold-fading) for 126 fifth grade students studying science-based materials over 6 weeks. Their results demonstrated the effectiveness of map-correction and scaffold-fading approaches in enhancing the summarization ability of students.

In summary, spatial strategies were shown to improve learning performance when compared to traditional approaches to instruction.

It was noted that most previous studies failed to differentiate the benefits of viewing author-provided graphics and those obtained through the construction of their own graphic. For example, Griffine et al. (1995) did not differentiate the base effects of constructing and viewing maps, because all of the students who received graphic organizers were also asked to construct their own.

Stull and Mayer (2007) had college students read a biology text. One group of students simultaneously viewed author-provided graphic organizers, another group constructed their own graphic organizers, and a third group studied without an additional task. These students then took retention and transfer tests. There were no significant differences between their retention scores. However, the transfer scores of students who viewed author-provided graphic organizers were better than those of the group of students who generated their own graphics and the control group.

When Stull and Mayer (2007) attempted to clarify whether the beneficial effects of using graphic learning strategies were due to viewing or doing maps, their results indicated that simply viewing author-provided graphics was more effective than both having students construct one's own graphic and traditional text learning methods.

In short, concept mapping strategies enhanced text comprehension and summarization abilities (Chang et al., 2002). Besides, Stull and Mayer (2007) research indicated that simply showing concept maps to students could effectively facilitate their comprehension and learning. Therefore, providing expert concept maps could be a very helpful strategy.

The second objective in this study is to employ concept maps in the proposed automatic summary assessment and feedback system, meanwhile investigating the effects of providing a concept map on the writing of summaries by students in the sixth grade.

1.4. Objective

First, training in the skills of summarization can be of tremendous help in reading comprehension, meanwhile several automatic summary assessment and feedback systems based on LSA have been developed to ease the work load of instructors with regard to grading summaries, and to provide immediate tailor-made feedback to students for improving their summary writing ability (Kintsch et al., 2000; Landauer et al., 2003; Landauer et al., 2009; Wade-Stein & Kintsch, 2004). However, previous studies have never clarified the question of whether more feedback concerning sentence relevance in student summaries significantly promotes students' summary writing performance. So the first objective in the study was to compare the effectiveness of relevance feedback on enhancing students' summary writing performance.

Second, other studies (Chang et al., 2002; Stull & Mayer, 2007) showed that providing expert concept maps could be a very helpful strategy for enhancing comprehension and summarization performance, thus the concept maps method was also employed in the proposed automatic summary assessment and feedback system. Therefore, the second objective in this study was investigating the effects of providing relevance information and concept maps on the writing of summaries by students in the sixth grade.

Finally, previous research (Bean & Steenwyk, 1984; Brown & Day, 1983) suggested that the ability to summarize information is a late developing skill, as many adult students failed to master summarization. However, many studies also showed that middle-grade students (10–14 years old) profited from instruction in concept maps and summarization (Chang et al., 2002; Griffin et al., 1995; Wade-Stein & Kintsch, 2004), so sixth graders were chosen for this study.

This study developed an automatic summary assessment and feedback system based on LSA that uses concept maps and concept words, and then investigated the effectiveness of providing relevance information and concept maps, and also examines the relative contribution of relevance feedback and concept map feedback on the writing of summaries by students in the sixth grade.

2. System development

The framework of the proposed system can be divided into a client side and a server side, as shown in Fig. 1. The client side includes an input interface and a feedback interface. The server side is responsible for the scoring of summaries, the evaluating of semantic similarity by comparing student summaries with a summary prepared by experts, and determines whether the student summaries employ appropriate concept words.

2.1. Client side

The input interface displays the articles to be read and provides a space for the writing of summaries for submission to the server. The feedback interface displays the feedback, including a score, semantic, and concept feedback.

Score feedback provides information about the length of the students' summary (character count), the overall score of the summary, and how well the content of each section of the source text has been covered. As shown in Fig. 2, clicking on the score feedback function button reveals this information on the top panel of the screen. The length and overall scores are indicated by numbers, while coverage rates for the sections are indicated by three horizontal bars.

Semantic feedback provides information related to semantic similarity between sentences in the user's summary and an expert summary. As shown in Fig. 3, clicking on the semantic feedback function button highlights highly-related sentences in the student's summary with bold type.

Concept feedback provides two types of information: a) A concept map created by experts to help users understand the conceptual structure of the source text, and b) The concept words extracted from the user's summary that are strongly

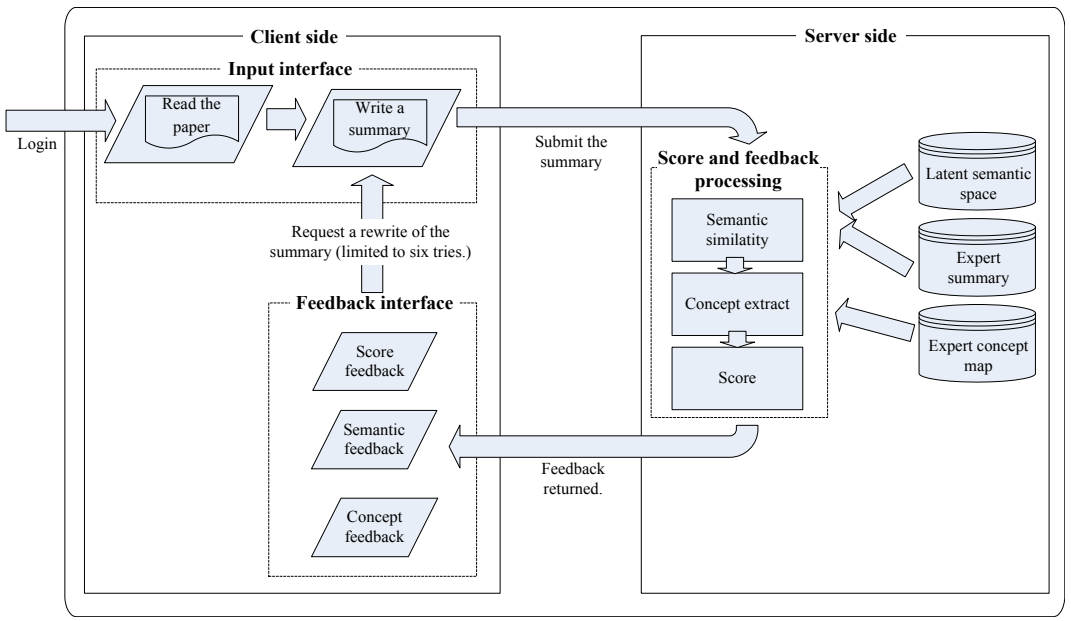

Fig. 1. The framework of the proposed system.

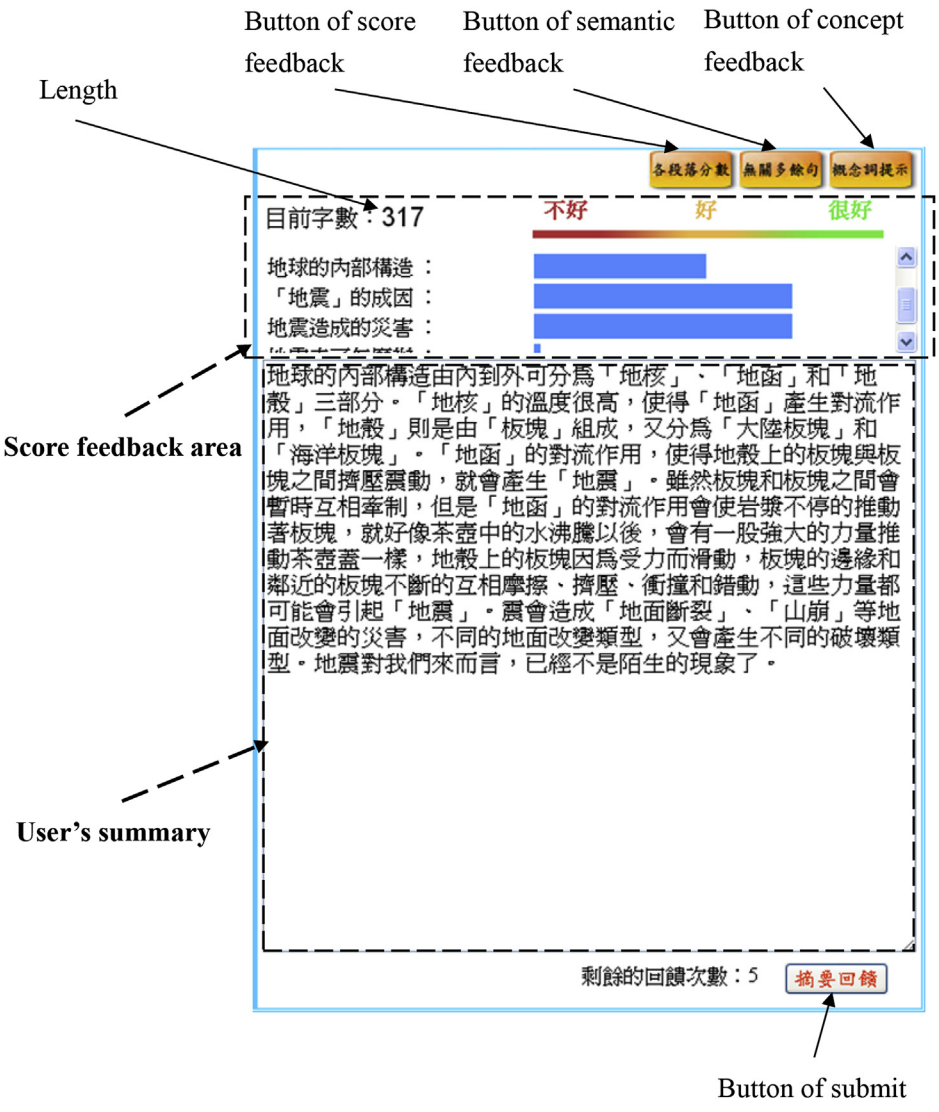

Fig. 2. Example of the score feedback.

relevant to the basic gist of the concept map. The box at the bottom of Fig. 4 contains the extracted content, with the concept words marked in blue. This reveals the important words in the sentence and whether they cover the main points.

2.2. Server side

The primary functions of the server include the evaluation of summaries and the provision of feedback. These processes require three databases containing the latent semantic space, the expert summaries, and the expert concept maps. In the following, we detail the databases and associated processes.

2.2.1. Latent semantic space

For the textual input, we used 262 Chinese articles obtained from science textbooks from the third to sixth grades. The system first delimited each sentence by punctuation, such as the Chinese period (。), Chinese exclamation point (！), Chinese semicolon (；), and Chinese question mark (？). Next, the system segmented each sentence into words, and tagged the parts of speech of words with a toolkit, named WECAn (Chang, Sung, & Lee, 2012; Sung, Chang, Lin, Hsieh, & Chang, in press). Finally, all words other than nouns and verbs were removed.

After the preprocessing, the system generated a 6393 × 262 term-to-document occurrence matrix, and then transformed the term-to-document matrix using SVD and dimension reduction methods into a latent semantic space with 250 dimensions.

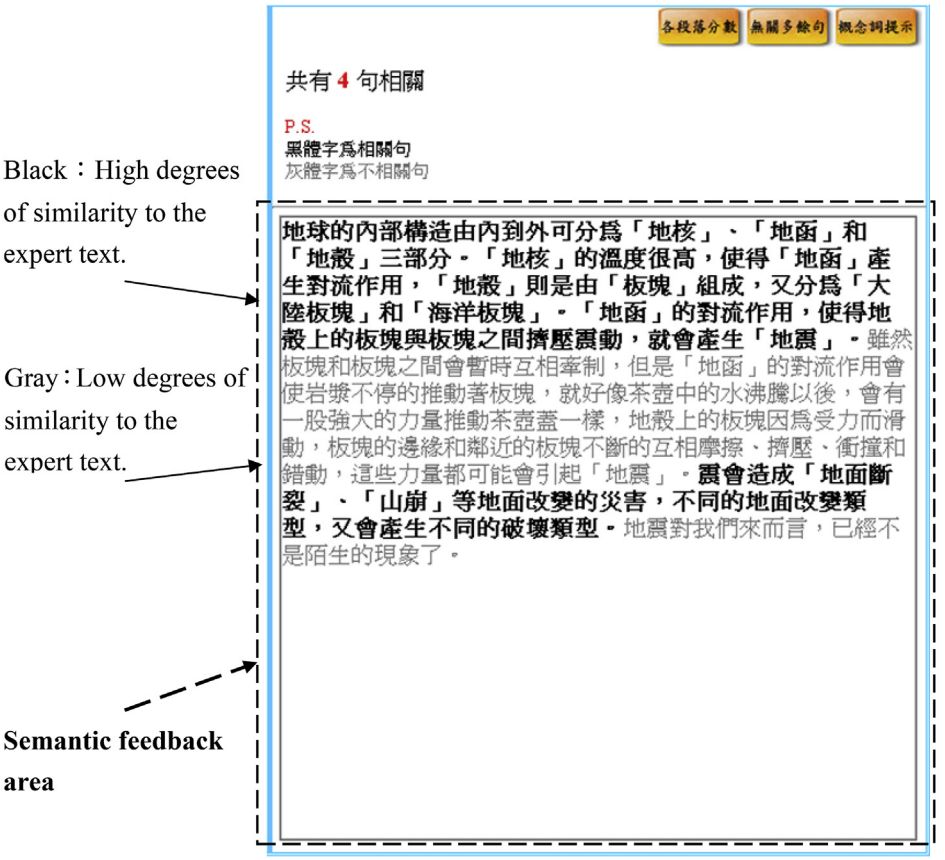

Fig. 3. Example of the semantic feedback.

Once the LSA space was constructed, it is easy to compute the semantic similarity between pairs of sentences. Each sentence, either from the source text, expert's summary, or student's summary, can be projected into the LSA space and to get a vector.

2.2.2. Expert summaries and concept maps

Three experts produced nine sets of summaries and concept maps (as shown in Fig. 4) for the corresponding nine source texts used in this research. The experts were two Chinese teachers and a PhD specialist in reading comprehension. Concept maps were drawn by the two Chinese teachers based on the source texts. The PhD specialist then joined the discussion to help to finalize the design of the summaries and concept maps.

The system also determined a threshold value for each sentence in the expert summaries. For determining the threshold value of sentences in the expert summaries, the system first calculated the similarities (the cosine values obtained via LSA) between sentences in the expert summary and those found in the corresponding source text. The threshold value of a given sentence in the expert summary was then derived by averaging all similarities between sentences in the corresponding source text and the given sentence in the expert summary. This information was then used as a reference for the calculation of semantic similarity and sentence relevance processes.

Meanwhile, the words and phrases used in the expert concept maps were also preprocessed, included word segmentation, part-of-speech tagging, and the removal of stop words. Only the nouns and verbs counted as concept words in the expert concept maps. After preprocessing, the information was then used as a reference for summary scoring and the extraction of concept terms from student summaries.

2.2.3. Calculation of semantic similarity and sentence relevancy

The system calculated the similarity between every sentence in each student's summary to those in the expert summary. Then, the similarity between every sentence in the student's summary and every sentence in the expert summary was calculated using LSA. These values were then compared to the threshold values of sentences in the expert summary to determine whether the sentences in the student's summary were relevant to the source text. This information about highly-related sentences is then used as a reference in the summary scoring process.

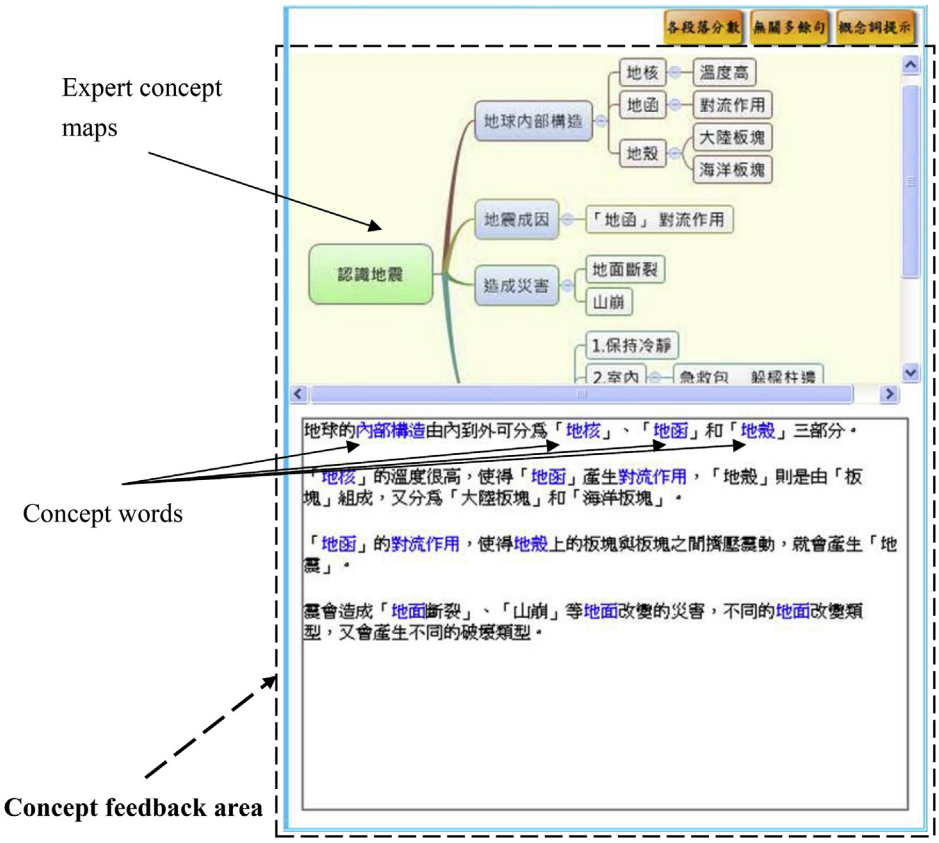

Fig. 4. Example of the concept feedback.

2.2.4. Extraction of concept terms

The concept words used in the expert concept maps were then identified by the system from the relevant sentences in the user's summary. This information is also used as a reference during the summary scoring process.

2.2.5. Summary scoring

The automatic summary scoring involves the integration of information obtained from the calculation of semantic similarity and the extraction of concept words to produce an overall score for the student's summary. This study modified the formulas proposed by Garner (1982) and Head et al. (1989) to include a semantic score, a concept word score, and a word count penalty. The relationship among these values is as follows:

$$\text{Summary score} = \frac{(\text{Semantic score} \times 70\%) + (\text{Concept word score} \times 30\%)}{\text{Penalty}}$$

The semantic score includes separate scores for each subtopic in which the numbers of sentences that are relevant to each subtopic are compiled and integrated into a single score. The formula is as follows:

$$\text{Semantic score} = \left(\sum \left(\frac{\text{Number of student sentences semantically relevant to subtopic}}{\text{Number of expert sentences regarding subtopic}} \times \frac{1}{\text{Number of all subtopic}} \right) \right) \times 100\%$$

The concept word score represents the ratio of the number of concept words used by the student divided by the number employed by the expert. The number of concept words used by the student is based only on the appearance of concept words identified in relevant sentences in the student summary. The number employed by the expert is based on the appearance of words in the entire expert concept map. The formula for the concept word score is as follows:

$$\text{Concept word score} = \left(\frac{\text{Number of concept words in relevant sentences in student summary}}{\text{Number of concept words in expert concept map}}\right) \times 100\%$$

Each summary has a word count limit and a penalty is imposed on summaries that exceed these limits. The penalty is based on the degree to which the user's summary exceeds the word count limit, the formula of which is as follows:

$$\text{Penalty} = \frac{\text{Actual word count user summary}}{\text{Word count limit}}$$

3. Methodology

3.1. Participants and design

154 sixth-grade students from 6 classes at an elementary school in New Taipei City were recruited for this study. The 154 students were randomly assigned to four experiment groups. The four groups contained 37, 37, 38, and 42 students, respectively. Because special needs students and students missing more than one practice session were removed from the analysis, 7, 7, 8, and 12 students' responses were not analyzed in the study for the four groups, respectively.

The study was conducted over a 6-week period for four groups of students: one week for pretests, four weeks of practice, and one week for posttests. The design involved two between-subject factors: semantic feedback (with, without) and concept feedback (with, without). Students in all groups received score feedback.

Two measures were used in the analysis: initial and final scores indicating improvements in the score of summary writing (test phase) as well as weekly scores indicating improvements in the score of summary writing (practice phase). Improvements in the scores of the test phase were calculated by subtracting the score of the pretest from the score of the posttest. Improvements in the scores of the practice phase were calculated by subtracting the score of the pretest from the weekly mean score.

3.2. Material

3.2.1. Source text

This study used nine source texts selected from science textbooks and rewritten by the researchers. The characters of these source texts were checked by three Chinese teachers and a PhD specialist in reading comprehension and found to be comparable on a variety of features, including style, length, difficulty, etc. All source texts were scientific texts written in an expository style, and so included a title, subtitle, and three or four topic sections. The length of the source texts ranged between 1200 and 1450 characters. The readabilities in Chinese Readability Index Explorer, CRIE 1.0 (Sung, Chen, Cha, Tseng, Chang, & Chang, 2015) of the source texts ranged between 6.42 and 8.06 grade ($M = 7.56$, $SD = .52$). Readability, as assessed by human experts, ranged from fifth grade to seventh grade ($M = 6.56$, $SD = .73$). The three language experts independently scored the readability of the source texts as well as five other characteristics, including narrativity, syntactic simplicity, word concreteness, referential cohesion, and causal cohesion (Graesser, McNamara, & Kulikowich, 2011) using a 9-point Likert scale. Inter-rater reliabilities between these three raters were all above .85. Any disagreements on readability was resolved by conferring among the raters.

3.2.2. Expert summaries and concept maps

As described in System Development Section.

3.3. Feedback system

As described in System Development Section.

3.4. Procedure

Every student received a pretest, practice sessions, and a posttest. The pretest, practice and posttest phases included one, seven, and one sessions, respectively. The only pretest session was held in the first week; practice sessions were held twice per week in the second to fourth weeks, and held only once in the fifth week. The only posttest session was held in the sixth week. The students practiced only a single text per session. In addition, the nine source texts were presented in the same order across the four experiment groups.

3.4.1. Pretest

The pretest phase included only one session in the first week. During the first 5 min, the experimenter explained the term “summarizing a text” and demonstrated the corresponding operational procedures for the computer-based system to all

participants. In the 35 min following the demonstration, all participants were provided an online text to summarize as a pretest. The students were able to compose, revise, and save their summaries as many times as they wished; however, the system did not provide any feedback during the pretest phase.

3.4.2. Practice phase

There were seven sessions in the practice phase. All students practiced only one article per session, and each session took approximately 40 min. In the first two sessions, the experimenter still provided instructions on the corresponding operational procedures before the students went to operate the system themselves. After logging in to their computer, the students then individually selected the predesignated text for that session. Afterwards, they could either compose their summary in the system or compose it in editing software and then copy-paste it into the text field. After submitting their summary, students could request score, semantic, or concept feedback, as appropriate to their group, by pressing the corresponding buttons in the upper-right corner of the screen (as shown in Figs. 2–4). The students could submit summaries and receive feedback up to six times in one session.

3.4.3. Posttest

All students took a posttest following completion of the practice phase. All procedures were the same as those used in the pretest, except that the online text was different.

3.5. Human scoring

3.5.1. Human raters

Three science teachers were recruited to rate the relative importance of each sentence in the source texts and then rate the students' summaries.

3.5.2. Rating the relative importance of the sentences in source texts

The three science teachers were asked to read a source text and then to rate how important each sentence would be to a summary of that source text as 1 (not suitable for its summary), 2 (suitable to appear in its summary), or 3 (must appear in its summary).

Those sentences rated as 3 were identified as the most important ideas in each source text. The number of important ideas of the nine source texts was about 12.78 per source text (SD = 3.67). The ratios of the number of important ideas of a source text divided by the number of sentences of the source text was about .38 (SD = .12).

The source texts were distributed to the raters over the course of three weeks, with a slightly increased load each week (2, 3, and 4 texts, respectively.) Each week, the raters independently rated the source texts and then resolved any disagreement in conference with the other raters. The initial disagreement rate was .75, .56, .32 for week 1, week 2 and week 3, respectively, showing that the rater reliability improved markedly over time.

3.5.3. Summary scoring

We modified the formulas proposed by Garner (1982) and Head et al. (1989) to include the proportion of key ideas and the word count penalty. The relationship among these values is as follows:

$$\text{Summary score} = \frac{\text{Proportion of important ideas}}{\text{Penalty}} \times 100\%$$

The proportion of important ideas refers to the ratio of the number of key ideas mentioned by the student divided by the number of important ideas of source text. The formula for the concept word score is as follows:

$$\text{Proportion of important ideas} = \frac{\text{Number of important ideas mentioned by students}}{\text{Number of important ideas of the source text}}$$

Each summary has a word count limit and a penalty is imposed on summaries that exceed these limits. The penalty is based on the degree to which the user's summary exceeds the word count limit, the formula of which is as follows:

$$\text{Penalty} = \frac{\text{Actual word count user summary}}{\text{Word count limit}}$$

The number of important ideas mentioned by students was the central focus in these formulae. The first scores calculated were the number of important ideas appearing in each student's summary. For training purposes, the three raters were asked to check 20 pretest summaries and 20 posttest summaries for the absence/presence of each important idea. We used the Kappa statistic as a measure of agreement between the three raters. The Kappa values were .864 (3 raters and 220 cases, 1 variable with 660 decisions in total) and .737 (3 raters and 260 cases, 1 variable with 780 decisions in total). These values

indicated that the raters had almost perfect and substantial agreement, respectively (Landis & Koch, 1977). The other 200 student summaries were then randomly assigned to the three raters for summary scoring.

4. Results

4.1. Correlations between human scoring and LSA-based automated scoring

The Pearson correlation between human scoring and LSA-based automated scoring was calculated to examine the validity of LSA-based automated scoring. Human scoring and automated scoring were moderately correlated ($r = .545$, $p < .01$).

4.2. Intervention effect

Two measures were used in the analysis: initial and final scores indicating improvements in the score of summary writing (test phase) as well as weekly scores indicating improvements in summary writing (practice phase). Improvements in the scores of the test phase were calculated by subtracting the score of the pretest from the score of the posttest. To calculate the improvement of the scores over the practice phase, the two session scores for each week were averaged together, yielding four mean scores for weeks 2–5. Then, the pretest score was subtracted from each of these weekly means, thus showing how much the various groups improved over time.

Table 1 gives the pretest scores, the mean improvements in the scores of test phase, the mean weekly improvements in scores throughout the practice phase for human scoring, and automated scoring. Due to resource limitations, the improvement scores for the practice phase were only calculated automatically.

4.2.1. Improvements in the summary writing score in test phase-human scoring

The human scoring score of the pretest was analyzed in a two-way between-subject analysis of variance (ANOVA) with semantic feedback (with or without), and concept feedback (with or without) as factors. The main effect of semantic feedback ($F(1, 116) = 1.803$, $MSE = 406.843$, $p = .182$, $\eta_p^2 = .015$) and concept feedback ($F(1, 116) = .881$, $MSE = 406.843$, $p = .350$, $\eta_p^2 = .008$) were not significant. Meanwhile the interaction between semantic feedback and concept feedback was also not significant ($F(1, 116) = 3.365$, $MSE = 406.843$, $p = .069$, $\eta_p^2 = .028$). This supported that the four groups had comparable beginning summarization performance.

Improvements in the scores of the test phase were analyzed in a two-way between-subject analysis of variance (ANOVA) with semantic feedback (with or without), and concept feedback (with or without) as factors. Only the effect of concept feedback ($F(1, 116) = 4.339$, $MSE = 526.240$, $p = .039$, $\eta_p^2 = .036$) was significant: The mean improvement in summary writing score for those in the test phase with the concept feedback condition ($M = 24.14$, $SD = 20.93$) was significantly higher than that of those with the without concept feedback condition ($M = 15.41$, $SD = 24.78$; As shown on Table 1 and Fig. 5(a)). All other main effects and interaction effects were not significant.

4.2.2. Improvements in summary writing score in test phase-automated scoring

The automated scoring score of the pretest was analyzed in a two-way between-subject analysis of variance (ANOVA) with semantic feedback (with or without), and concept feedback (with or without) as factors. The effect of semantic feedback ($F(1, 116) = 12.201$, $MSE = 348.664$, $p = .001$, $\eta_p^2 = .095$) was significant, and the two-way interaction between semantic feedback and concept feedback were both significant ($F(1, 116) = 14.158$, $MSE = 348.664$, $p < .001$, $\eta_p^2 = .109$). This indicated that the four groups did not have comparable beginning summarization performance.

So the improvements in the automated scoring scores of the test phase were then analyzed in a two-way between-subject analysis of covariance (ANCOVA) with semantic feedback (with or without), and concept feedback (with or without) as between-subject factors, and the pretest score as covariate. Only the effect of concept feedback ($F(1, 115) = 3.874$, $MSE = 222.162$, $p = .051$, $\eta_p^2 = .033$) was marginally significant: The mean improvement in summary writing score of those in the test phase with the concept feedback condition ($M' = 24.92$) was significantly higher than that of those with the without concept feedback condition ($M' = 19.55$; As shown on Table 1 and Fig. 5(b)).

4.2.3. Weekly improvements in scores throughout practice phase

Because the beginning summarization performance in the automated scoring of the pretest wasn't comparable, the weekly improvement in the scores of summary writing throughout the practice phase was analyzed using the three-way between-subject and within-subject mixed ANCOVA with semantic feedback (with or without), and concept feedback (with or without) as the between-subject factors, study duration (week 1/2/3/4) as the within-subject factor, and the pretest score as the covariate.

The main effect of study duration ($F(3, 339) = 20.096$, $MSE = 167.325$, $p < .001$, $\eta_p^2 = .151$) was significant. The Fisher's protected LSD test revealed that the improvements observed in week 1 ($M' = 11.33$) and week 2 ($M' = 15.03$) were not significantly different; however, they were both significantly lower than the improvements observed in week 3 ($M' = 30.30$) and week 4 ($M' = 44.23$). In addition, the improvement in the scores in week 3 was also significantly lower than that of week 4 at $p < .05$.

Table 1
Mean pretest scores, mean improvement in the scores of test phase, mean improvement in weekly scores and standard deviation, as a function of semantic feedback, concept feedback, and period of study.

Semantic feedback	Concept feedback	Pretest		Improvement in the scores of test phase			Improvement in weekly scores in practice phase											
							Week 1			Week 2			Week 3			Week 4		
		M	SD	M	SD	*M′*	M	SD	*M′*	M	SD	*M′*	M	SD	*M′*	M	SD	*M′*
Human scoring	Without	63.72	20.51	13.42	20.96	–	–	–	–	–	–	–	–	–	–	–	–	–
	With	53.50	17.31	27.79	18.50	–	–	–	–	–	–	–	–	–	–	–	–	–
	Without	52.02	21.26	17.40	28.32	–	–	–	–	–	–	–	–	–	–	–	–	–
	With	55.31	21.33	20.48	22.83	–	–	–	–	–	–	–	–	–	–	–	–	–
Automated scoring	Without	49.76	19.57	7.31	23.48	17.31	29.30	25.68	12.71	5.70	24.10	18.09	16.27	24.36	28.17	29.68	18.82	38.97
	With	34.12	18.90	26.98	19.78	25.63	17.66	26.70	16.40	20.44	23.14	19.17	34.14	20.84	32.93	47.39	18.29	46.44
	Without	25.03	16.00	29.74	19.17	21.79	16.52	23.42	6.66	19.48	23.38	9.64	38.12	15.79	28.67	52.73	19.81	45.35
	With	35.04	19.95	24.90	17.48	24.21	11.23	18.44	9.56	14.88	21.77	13.21	33.05	18.46	31.45	47.42	16.15	46.17

As shown on Table 1 and Fig. 6, the two-way interaction between semantic feedback and study duration was also significant (F (3, 339) = 3.762, MSE = 167.325 p = .011, η_p^2 = .032).

Regarding the weekly improvement scores between four kinds of study duration in the same treatment, the simple main effect analysis indicated that for those in the without semantic feedback condition, the improvements observed in week 1 (M' = 8.98) and week 2 (M' = 13.07) were not significantly different; however, they were both significantly lower than the improvements observed in week 3 (M' = 25.20) and week 4 (M' = 38.53). In addition, the improvement in the scores in week 3 was also significantly lower than that of week 4 (F(3, 339) = 5.854, MSE = 167.325 p = .001, η_p^2 = .049). For those in the with semantic feedback condition, the improvements observed in week 1 (M' = 13.88) and week 2 (M' = 17.18) were not significantly different; however, they were both significantly lower than the improvements observed in week 3 (M' = 35.59) and week 4 (M' = 50.07). In addition, the improvement in the scores in week 3 was also significantly lower than that of week 4 (F (3, 339) = 23.266, MSE = 167.325 p < .001, η_p^2 = .170).

Regarding the weekly improvement scores between the with and without semantic feedback conditions with the same study duration, the simple main effect analysis indicated that in the first week of the practice phase, the improvement in the scores of those in the with the semantic feedback condition (M' = 8.15) was significantly lower than that of those in the with the without semantic feedback condition (M' = 14.15; F (1, 452) = 4.134, MSE = 239.376, p = .043, η_p^2 = .009); In the second week of the practice phase, the improvement in the scores of those in the with the semantic feedback condition (M' = 11.96) was also significantly lower than that of those in the with the without semantic feedback condition (M' = 18.07; F(1, 452) = 4.271, MSE = 239.376, p = .043, η_p^2 = .009); while, in the third week of the practice phase, the improvement in the scores of those in the with semantic feedback condition (M' = 30.23) was equal to those in the with the without semantic feedback condition (M' = 30.18; F (1, 452) = .000, MSE = 239.376, p = .987, η_p^2 < .001); also, in the fourth week of the practice phase, the improvement in the scores of those in the with the semantic feedback condition (M' = 45.51) was equal to those in the with the without semantic feedback condition (M' = 42.95; F (1, 452) = .725, MSE = 239.376, p = .395, η_p^2 < .001).

4.3. Submission count and tool use

Two measures were used in the analysis: submission count during the test phase as well as submission count throughout the practice phase.

As mentioned in the Method section, in the pre- and post-test sessions, the students could compose, revise, and save their summaries as many times as they wished without receiving any feedback during the test phase. In the practice phase, the students could compose, revise and submit summaries up to six times in one session.

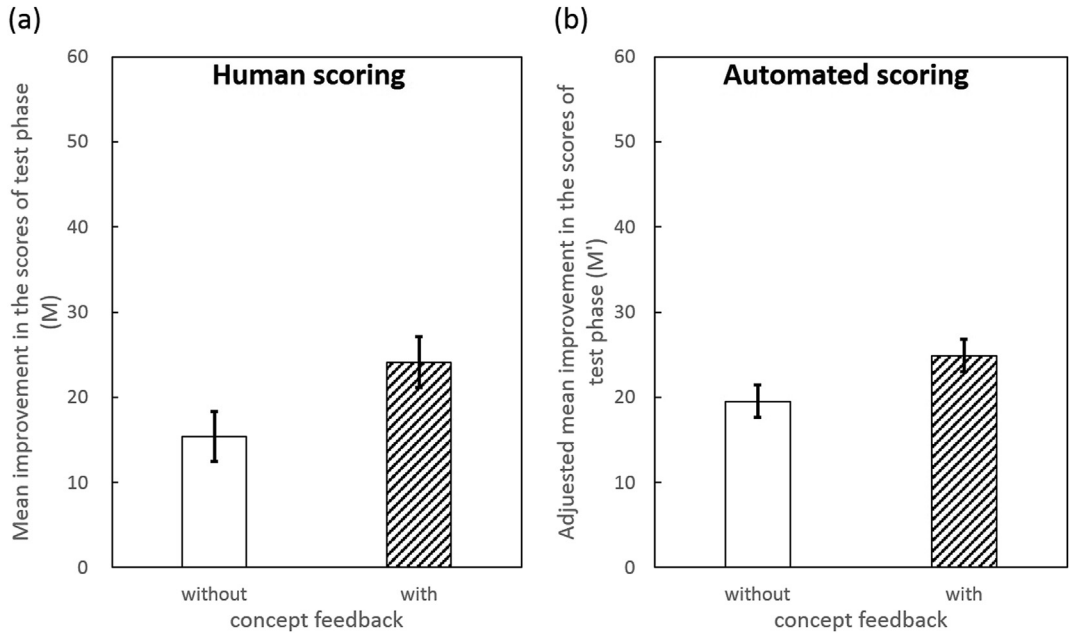

Fig. 5. (a) Mean improvement in the scores of the without (empty) and with (solid) concept feedback conditions for the testing phase. Error bars represent standard errors. (b) Adjusted mean improvement in the automated scores of the test phase for the without (empty) and with (solid) concept feedback conditions. Error bars represent standard errors.

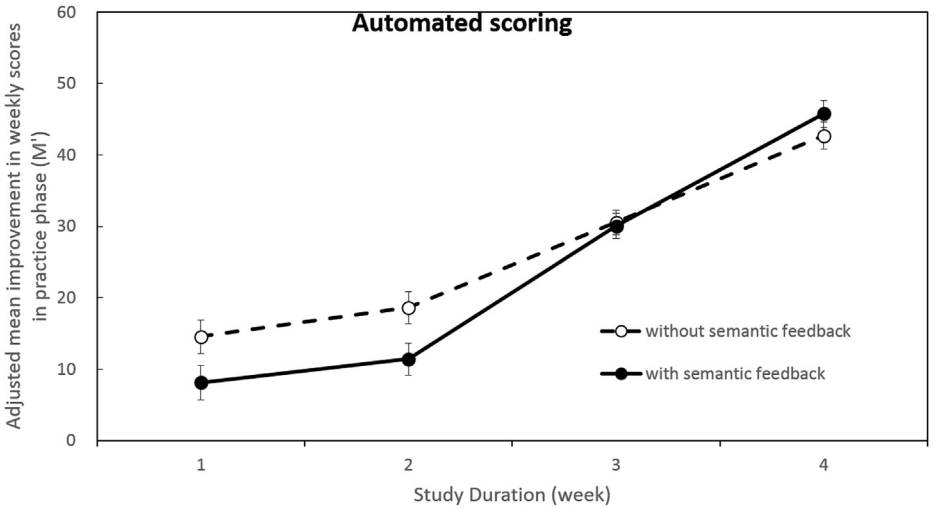

Fig. 6. The interaction between the adjusted mean improvement in weekly scores for the without (○) and with (●) semantic feedback conditions and study duration during the practice phase. Error bars represent standard errors.

After each submission, the students could request and receive a variety of feedback, as appropriate to their group, until they either composed a new version of the summary or logged out of the system. After submitting their summary, students could receive score, semantic, or concept feedback, as appropriate to their group. The submission count for each session was automatically recorded by the system. Because each additional revision and subsequent submission necessitates increased usage, we used submission count as an indicator of amount of tool use.

4.3.1. Submission count during the test phase

Submission count during the test phase was analyzed in a three-way between-subject and within-subject mixed ANOVA with semantic feedback (with or without), concept feedback (with or without) as the between-subject factors, and session (pretest or posttest) as the within-subject factor. Only the effect of session ($F(1, 116) = 7.747$, $MSE = 2.711$, $p = .006$, $\eta_p^2 = .063$) was significant: The submission count in the posttest session (M = 3.09, SD = 1.32) was significantly less than that of those in

the pretest session (M = 3.68, SD = 2.04). The main effect of semantic feedback ($F(1, 116) = 2.862$, $MSE = 3.216$, $p = .093$, $\eta_p^2 = .024$) and concept feedback ($F(1, 116) = .374$, $MSE = 3.216$, $p = .542$, $\eta_p^2 = .003$) were not significant. Meanwhile the interaction between semantic feedback and concept feedback was also not significant ($F(1, 116) = .971$, $MSE = 3.216$, $p = .971$, $\eta_p^2 < .001$). All other main effects and interaction effects were not significant, as shown in Fig. 7.

4.3.2. Submission throughout practice phase

Submission count throughout the practice phase was analyzed in a three-way between-subject and within-subject mixed ANOVA with semantic feedback (with or without), and concept feedback (with or without) as the between-subject factors, and study duration (week 1/2/3/4) as the within-subject factor.

The main effect of study duration ($F(3, 348) = 11.609$, $MSE = .869$, $p < .001$, $\eta_p^2 = .091$) was significant. The Fisher's protected LSD test revealed that the submission count in week 1 (M = 4.43, SD = 1.00) was significantly less than the submission counts in all other weeks. The submission count in week 3 (M = 4.80, SD = 1.20) and week 4 (M = 4.83, SD = 1.42) were not significantly different; however, they were both significantly less than the submission counts in week 2 (M = 5.13, SD = .98) at $p < .05$.

The two-way interaction between concept feedback and study duration was also significant ($F(3, 348) = 2.905$, $MSE = .869$, $p = .035$, $\eta_p^2 = .024$).

Regarding the submission count between the four kinds of study duration in the same treatment, the simple main effect analysis indicated that for those in the without concept feedback condition, the submission count in week 2 (M = 5.20, SD = .94) was significantly great than the submission counts in all other conditions. Meanwhile, the submission count in week 1(M = 4.34, SD = .96), week 3 (M = 4.54, SD = 1.24), and week 4 (M = 4.62, SD = 1.51) were not significantly different from each other ($F(3, 348) = 9.385$, $MSE = .869$, $p < .001$, $\eta_p^2 = .075$). For those in the with concept feedback condition, the submission count in week 1 (M = 4.51, SD = 1.04) was significantly less than the submission counts in all other conditions. Meanwhile, the submission count in week 2 (M = 5.07, SD = 1.03), week 3 (M = 5.06, SD = 1.10), and week 4 (M = 5.03, SD = 1.31) were not significantly different from each other ($F(3, 348) = 5.129$, $MSE = .869$, $p = .002$, $\eta_p^2 = .042$).

Regarding the submission count between those in the with and without concept feedback conditions within the same study duration, the simple main effect analysis indicated that in the first week of the practice phase, the submission count of those in the without concept feedback condition (M = 4.34, SD = .96) were equal to those in the with concept feedback condition (M = 4.51, SD = 1.04; $F(1, 464) = .628$, $MSE = 1.328$, $p = .43$, $\eta_p^2 = .001$); also, in the second week of the practice phase, the submission count of those without the concept feedback condition (M = 5.20, SD = .94) was equal to those with the without semantic feedback condition (M = 5.07, SD = 1.03; $F(1, 464) = .402$, $MSE = 1.328$, $p = .53$, $\eta_p^2 = .001$); while, in the third week of the practice phase, the submission count of those in the without concept feedback condition (M = 4.54,

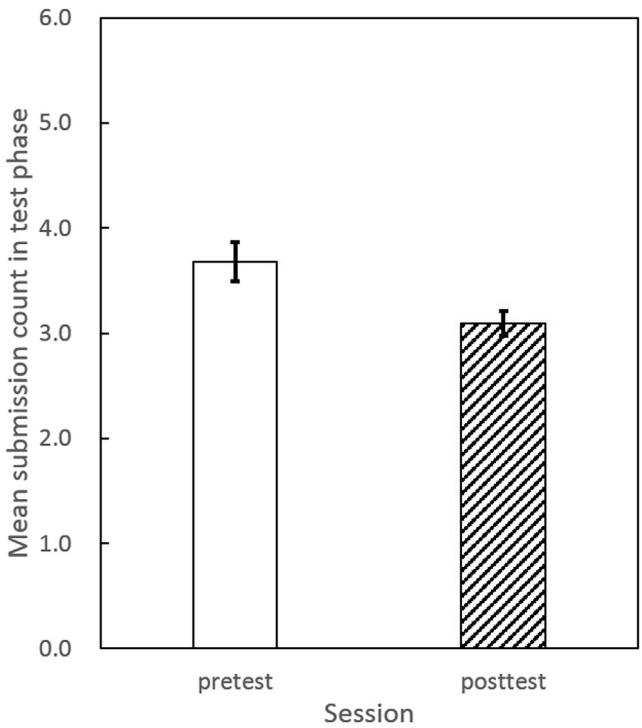

Fig. 7. Mean submission counts in the pretest and posttest sessions of the testing phase. Error bars represent standard errors.

SD = 1.24) was significantly less than that of those in the with concept feedback condition (M = 5.06, SD = 1.10; F (1, 464) = 6.031, MSE = 1.328, p = .02, η_p^2 = .013); also, in the fourth week of the practice phase, the submission count of those in the without concept feedback condition (M = 4.62, SD = 1.51) was significantly less than that of those in the with concept feedback condition (M = 5.03, SD = 1.31; F (1, 464) = 3.922, MSE = 1.328, p = .048, η_p^2 = .008).

In summary, the submission count increased from the first week to the second week of the practice phase for students regardless of condition. Then, the submission count maintained the same level during the following two weeks for students who received concept feedback. For those who didn't receive concept feedback, the submission count decreased in the third and fourth weeks. A summary of these results is shown in Fig. 8.

5. Discussion and conclusion

5.1. Empirical findings

This study developed a means of assessing the quality of text summaries, with the automatic provision of feedback related to overall scoring as well as semantic feedback, concept maps, and concept words, depending on experimental group. Our results led to the following conclusions:

First, human scoring and automated scoring were moderately correlated. Additionally, whether scored automatically or by humans, the results consistently showed that only concept feedback significantly affects improvements in summary writing scores in the test phase. This further supports the notion that automated scoring is comparable to human scoring.

Second, the significant improvements observed in the weekly scores throughout the practice phase demonstrated the effectiveness of the proposed system in improving the summary writing skills of students. The proposed system increases the opportunities for students to engage in improving their work through the provision of immediate feedback.

In addition, the significantly decreasing submission count in the posttest session showed that the students performed better, in both human scoring scores and automated scoring scores, with fewer revisions and no feedback in the posttest session. This phenomena supports the idea that the students were learning to master summary writing skills, rather than learning to rely on the support tools. Their increased skill let them attain a satisfactory result with fewer revisions.

Third, although the effect of concept feedback on the weekly improvements in scores throughout practice phase wasn't significant, we could still see the trend that the performance of the students with concept feedback outperformed the students without concept feedback across all conditions in the practice phase (see the M′ column on Table 1). Additionally, despite the non-significant result in the practice phase, the testing phase did show a significant improvement overall. This implies that concept feedback still benefited students during the practice phase, and that this benefit somehow led to higher scores on the posttest.

In the practice phase, the without concept feedback group's submission count significantly decreased to be less than those students in the with concept feedback group during the third and fourth weeks of the practice phase; meanwhile during the test phase their submission count didn't significantly differ from the with concept feedback group. It's reasonable, then, to infer that during the practice phase because the without concept feedback group made less use of the feedback system they learned less than the students in the with concept feedback group. Furthermore, the different effectiveness between the with and without concept feedback conditions didn't result from the submission count in the test phase itself because there was no significant difference between the submission counts of the two groups in that phase.

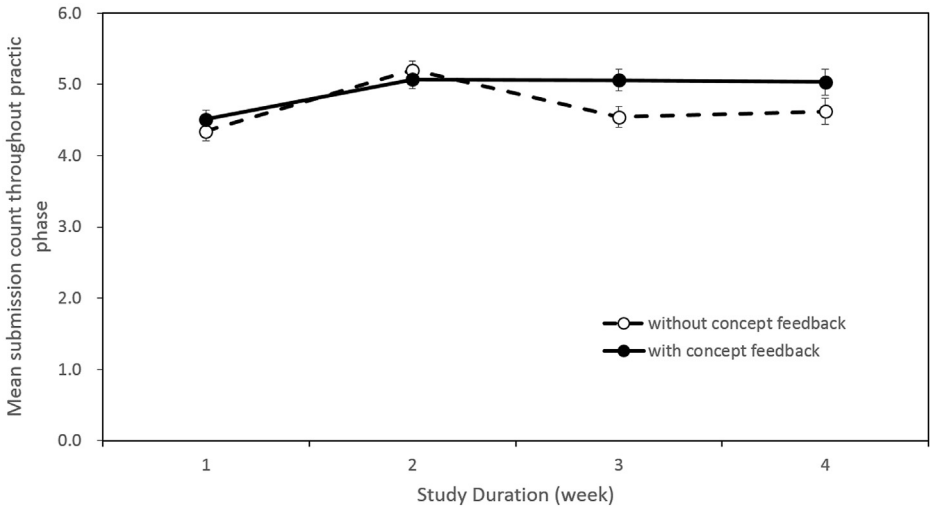

Fig. 8. The interaction between mean submission counts throughout the practice phase for the without (○) or with (●)concept feedback conditions and study duration. Error bars represent standard errors.

Fourth, the significant interaction effect between study duration and semantic feedback in weekly improvements in scores throughout the practice phase revealed that the students with semantic feedback initially needed more time to become familiar with the function before embarking on the program in earnest; however they progressed faster and caught up with the performance of the other students in the third week of the practice phase. Because the improvement rate was higher, it is possible that had there been a fifth week of practice the semantic feedback groups would have surpassed the without feedback groups, and such a plausibility is further discussed in Section 5.3.1.

5.2. Practical implications

5.2.1. Practice with feedback is helpful

Unlike previous researchers (Franzke et al., 2005; He, Hui, & Quan, 2009; Wade-Stein & Kintsch, 2004), the present research moved beyond the improvements observed between pre- and post-tests to examine the effect of the proposed feedback system in improving learning process in an on-going manner. During training (see Table 1 and Fig. 6), summary writing performance showed progressive improvements throughout the practice phase. The phenomenon supported that practicing summary writing with feedback did help with learning summarization.

No significant improvements were observed from the first to the second week, indicating that students may require time to become familiar with the program. For future curriculums or tutor systems, we recommend adding one or two weeks to familiarize students with the new technique.

5.2.2. Effectiveness of concept feedback

Although the apparent trend that students of the concept feedback groups outperformed the students of the other groups during the practice phase was not statistically significant, concept feedback was shown to have a positive, significant effect on learning performance in the test phase, which suggests that providing students with the concept structure of the source text can help to improve the summary writing skills of students. Practice without concept feedback is insufficient.

It can be noted that while the concept feedback groups ended up with superior scores in the testing phase, their rate of improvement in the practice phase started out relatively flat, only to pick up in later weeks. This result is consistent with that of McCagg and Dansereau (1991), who indicated that several obstacles must be overcome before the knowledge-map technique can be successfully implemented. Students must be given opportunities to be instructed in the technique, practice reading maps prepared by experts, and practice making their own maps before attempting assignments. Thus, we recommend adding lectures regarding the preparation and use of spatial learning strategy, such as concept map and extending the practice phase to provide students sufficient practice reading well-conceived maps in order to grasp the meaning contained within them.

5.3. Limitations and further work

5.3.1. Semantic feedback effect and study duration

The significant interaction effect between study duration and semantic feedback in the practice phase showed that the students in the with semantic feedback group started out with lower scores but progressed faster than the students of the without semantic feedback group, such that their performance caught up with the others in the third and fourth weeks. Moreover, we observed the trend that the improvement in scores in the fourth week of the students of the with semantic feedback group was higher than that of the students of the without semantic feedback group, although this was not statistically significant. There are many potential explanations for this initial slump followed by a superior improvement rate. One possibility is that students could have taken some time to acclimate to the system before being able to skillfully utilize its potential. Further research with longer practice phases could clarify the effect of semantic feedback on the summary writing skill of students.

5.3.2. Redundancy effect

According to the multimedia principle (Clark & Mayer, 2011; Liu, Lin, Gao, Teh, & Kalyuga, 2015), providing both words and graphics is more conductive to learning compared to just text or graphics alone. In the present research, the semantic and concept feedback provided the learners different information in different formats, and did not only highlight the inclusion of relevant vs. irrelevant information. First, semantic feedback provides feedback on the relevancy of sentences in a student's summary, while the concept feedback provides feedback not only on the relevancy of words in student's summary but also on the concept/knowledge structure of the original text (the concept map). Second, the semantic feedback only provides feedback on what the student wrote, and does not hint at what information the student is missing, whereas the concept feedback provides the whole conceptual structure of the original text no matter what answer is submitted. Third, the semantic feedback is textual (words) in nature, whereas the concept feedback provides the conceptual structural of the original text in a graphical format (picture). So providing both kinds of information could be a reasonable arrangement.

However, learners' performance isn't always improved as more kinds of information is provided, and in fact more information can sometimes be worse. This phenomenon is called the redundancy effect. According to Bobis, Sweller, and Cooper (1993, p. 2) “The redundancy effect occurs when students must process material intended to be informative but which is in fact redundant to other presented material.” and Yeung, Jin, and Sweller (1997, p. 3), “the redundancy effect occurs when the

learner is required to process nonessential information. It is this processing of unnecessary information that imposes an undue cognitive load." as well as the definition of the redundancy effect in the Encyclopedia of the Sciences of Learning (Jin, 2012, pp. 2787–2788), "The *redundancy effect* refers to the phenomenon in instruction where learning is hindered when additional information is presented to learners compared to the presentation of less information." for information to count as redundant information, it must be both nonessential and hinder students' learning.

The results of the test phase, both for human scoring scores and automated scoring scores, only found a significant effect for concept feedback, so there was no supportive evidence showing that including both kinds of feedback hindered learning. Meanwhile, the results of the practice phase also only showed a significant semantic feedback effect and a significant interaction between semantic and study duration, so once again there seemed to be no disadvantage to including all forms of feedback. Thus there was no sufficient reason to infer that a redundancy effect was found in this study.

Further research could add new functions to record the response time of students for reading feedback modules and writing tasks, as well as include additional measurement tools for observing how much feedback was provided to students. This could help teachers to better understand the difficulties faced by students and thereby enable instructors to select appropriate solutions to facilitate their development.

5.3.3. Experimental and statistical controls

Both experimental control methods and statistical methods can reduce error variance and obtain accurate estimates of treatment effects. In the experimental control approach, researchers usually randomize subject to various treatment levels. In the statistical approach, the analysis of covariance (ANCOVA) procedure is frequently applied to remove the biases caused by factors that are difficult or impossible to control by experimental methods.

However there are many assumptions that should be satisfied before applying the ANCOVA procedure, such as the independence, normality and homogeneity of the variances of the residuals, the linearity of regression, and the homogeneity of regression slopes (Braver, MacKinnon, & Page, 2003; Hick & Turner, 1999; Kirk, 1995). Although these assumptions were met in the present research, s-curves, dosage curves, or threshold effects are widespread in real situations. These kinds of situations are difficult to prevent or resolve.

Further research should first consider using as many kinds of experimental control methods as possible, and then consider applying additional statistical control methods to remove any remaining biases. For example, homogenously assigning subjects to each experiment group according to their pretest performance, rather than simply randomizing subjects to various treatment levels, could better prevent the pre-test differences problem we encountered.

Acknowledgement

The authors appreciate the funding support from the Ministry of Science and Technology, Republic of China (No: MOST 104-2511-S-003 -017 -MY3; MOST 104-2511-S-003 -012 -MY3; NSC 102-2911-I-003 -301).

References

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science, 33*(5–6), 381–412.

Bean, T. W., Singer, H., Sorter, J., & Frazee, C. (1986). The effect of metacognitive instruction in outlining and graphic organizer construction on students' comprehension in a tenth-grade world history class. *Journal of Reading Behavior, 18*(2), 153–169.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth-graders' summary writing and comprehension. *Journal of Reading Behavior, 16*(4), 287–306.

Block, C. C., & Pressley, M. (2002). In *Comprehension instruction: Research-based best practices*. New York: The Guilford Press.

Bobis, J., Sweller, J., & Cooper, M. (1993). Cognitive load effects in a primary-school geometry task. *Learning & Instruction, 3*(1), 1–21.

Braver, S. L., MacKinnon, D. P., & Page, M. (2003). Analysis of covariance. In *Levine's guide to SPSS for analysis of variance* (2nd ed., pp. 120–131). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*(1), 1–14.

Brunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106*(3), 849–858.

Chang, K. E., Sung, Y. T., & Chen, I. D. (2002). The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education, 71*(1), 5–23.

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012, November). A Chinese word segmentation and POS tagging system for readability research. In *Paper presented at 42nd annual meeting of the society for computers in psychology (SCiP 2012), Minneapolis, MN*.

Clark, R. C., & Mayer, R. E. (2011). *E-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning* (3rd ed.). San Francisco, CA: John Wiley & Sons.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIS, 41*(6), 391–407.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2–3), 285–307.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: computer support for comprehension and writing. *Journal of Educational Computing Research, 33*(1), 53–80.

Garner, R. (1982). Efficient text summarization: costs and benefits. *The Journal of Educational Research, 75*(5), 275–279.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

Griffin, C. C., Malone, L. D., & Kameenui, E. J. (1995). Effects of graphic organizer instruction on fifth-grade students. *The Journal of Educational Research, 89*(2), 98–107.

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction, 28*(4), 1–11.

He, Y., Hui, S. C., & Quan, T. T. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers & Education, 53*(3), 890–899.

Hick, C. R., & Turner, K. V. (1999). Miscellaneous topics. In *Fundamental concepts in the design of experiments* (5th ed., pp. 442–492). New York, NY: Oxford University Press.
Jin, P. (2012). Redundancy effect. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (2012 ed.) Retrieved from http://link.springer.com/referencework/10.1007/978-1-4419-1428-6/page/162.
Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7*(3), 161–195.
Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R., & the LSA Research Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8*(2), 87–109.
Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363.
Kirk, R. E. (1995). Analysis of covariance. In *Experimental design procedures for the behavioral science* (3rd ed., pp. 706–750). Pacific Grove, CA: Brooks/Cole.
Lan, Y. J. (2015). Contextual EFL learning in a 3D virtual environment. *Language Learning & Technology, 19*(2), 16–31.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.
Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295–308.
Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice, 48*(1), 44–52.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Lenhard, W., Baier, H., Endlich, D., Schneider, W., & Hoffmann, J. (2013). Rethinking strategy instruction: direct reading strategy instruction versus computer-based guided practice. *Journal of Research in Reading, 36*(2), 223–240.
Liu, T. C., Lin, Y. C., Gao, Y., Yeh, S. C., & Kalyuga, S. (2015). Does the redundancy effect exist in electronic slideshow assisted lecturing? *Computers & Education, 88*, 303–314.
Malone, L. D., & Mastropieri, M. A. (1992). Reading comprehension instruction: summarization and self-monitoring training for students with learning disabilities. *Exceptional Children, 58*(3), 270–279.
McCagg, E. C., & Dansereau, D. F. (1991). A convergent paradigm for examining knowledge mapping as a learning strategy. *The Journal of Educational Research, 84*(6), 317–324.
Stull, A. T., & Mayer, R. E. (2007). Learning by doing versus learning by viewing: three experimental comparisons of learner-generated versus author-provided graphic organizers. *Journal of Educational Psychology, 99*(4), 808–820.
Sung, Y. T. , Chang, T. H., Lin, W. C., Hsieh, K. S., & Chang, K. E. CRIE: An automated analyzer for Chinese texts. *Behavior Research Method*. (in press). http://dx.doi.org/10.3758/s13428-015-0649-1.
Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H., & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods, 47*(2), 340–354.
Sung, Y. T., Wu, M. D., Chen, C. K., & Chang, K. E. (2015). Examining the online reading behavior and performance of fifth-graders: evidence from eye-movement data. *Frontiers in Psychology, 6*, 665.
Wade-Stein, D., & Kintsch, E. (2004). Summary street: interactive computer support for writing. *Cognition and Instruction, 22*(3), 333–362.
Weisberg, R., & Balajthy, E. (1990). Development of disabled readers' metacomprehension ability through summarization training using expository text: results of three studies. *Journal of Reading, Writing, and Learning Disabilities International, 6*(2), 117–136.
Yeung, A. S., Jin, P., & Sweller, J. (1997). Cognitive load and learner expertise: split-attention and redundancy effects in reading with explanatory notes. *Contemporary Educational Psychology, 23*(1), 1–21.
Zimmerman, B. J. (1998). Academic studying and the development of personal skill: a self-regulatory perspective. *Educational Psychologist, 33*(2–3), 73–86.
Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: an overview and analysis. In B. J. Zimmerman, & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives: Mahwah, NJ*. Lawrence Erlbaum Associates.
Zimmerman, B. J. (2002a). Achieving academic excellence: a self-regulatory perspective. In M. Ferrari (Ed.), *The pursuit of excellence through education* (pp. 85–110). Mahwah, NJ: Lawrence Erlbaum Associates.
Zimmerman, B. J. (2002b). Becoming a self-regulated learner: an overview. *Theory into Practice, 41*(2), 64–70.