Behav Res (2018) 50:2144–2161
https://doi.org/10.3758/s13428-017-0982-7

Computerized summary scoring: crowdsourcing-based latent semantic analysis

Haiying Li[1] · **Zhiqiang Cai**[2] · **Arthur C. Graesser**[2,3]

Published online: 3 November 2017

Abstract In this study we developed and evaluated a crowdsourcing-based latent semantic analysis (LSA) approach to computerized summary scoring (CSS). LSA is a frequently used mathematical component in CSS, where LSA similarity represents the extent to which the to-be-graded target summary is similar to a model summary or a set of exemplar summaries. Researchers have proposed different formulations of the model summary in previous studies, such as pregraded summaries, expert-generated summaries, or source texts. The former two methods, however, require substantial human time, effort, and costs in order to either grade or generate summaries. Using source texts does not require human effort, but it also does not predict human summary scores well. With human summary scores as the gold standard, in this study we evaluated the crowdsourcing LSA method by comparing it with seven other LSA methods that used sets of summaries from different sources (either experts or crowdsourced) of differing quality, along with source texts. Results showed that crowdsourcing LSA predicted human summary scores as well as expert-good and crowdsourcing-good summaries, and better than the other methods. A series of analyses with different numbers of crowdsourcing summaries demonstrated that the number (from 10 to 100) did not significantly affect performance. These findings imply that crowdsourcing LSA is a promising approach to CSS, because it saves human effort in generating the model summary while still yielding comparable performance. This approach to small-scale CSS provides a practical solution for instructors in courses, and also advances research on automated assessments in which student responses are expected to semantically converge on subject matter content.

Keywords LSA similarity · Crowdsourcing · Computerized summary scoring

In this study we developed and evaluated a crowdsourcing-based approach to automatically scoring students' summaries, called *crowdsourcing-based latent semantic analysis* (LSA; Li, Cai, & Graesser, 2016; hereafter referred to as *crowdsourcing LSA*). We compared this crowdsourcing LSA method with seven other LSA methods based on sets of summaries from different sources (experts or crowdsourced) of differing quality (good, intermediate, and poor), along with source texts. The crowdsourcing LSA similarity scores were computed by comparing a target student summary with all other students' summaries that were to be graded (hereafter called a *crowdsourcing model summary*). We hypothesized that crowdsourcing LSA would predict human summary scores as well as, if not better than, the other LSA methods. The advantage of using crowdsourcing summaries as a model summary is that it maximally represents the content of students' written summaries, in contrast to using one source text, one expert summary, or several expert summaries, none of which reflect the actual student content. This study contributes to research on computerized summary scoring (CSS) by proposing an approach that does not require human effort to either

✉ Haiying Li
haiying.li@gse.rutgers.edu

[1] Graduate School of Education, Rutgers - The State University of New Jersey, New Brunswick, NJ 08901, USA

[2] Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA

[3] Department of Psychology, University of Memphis, Memphis, TN, USA

write ideal summaries or pregrade summaries. Therefore, this method saves instructors additional time, effort, and costs, and is a practical solution for instructors who teach courses both in traditional classroom settings and online.

The following introductory section elaborates on the rationale for this study, by reviewing related literature. The review starts with the complex cognitive activities that occur during the processes of summary writing, and it specifies their important roles in deep reading comprehension and learning. The current CSS techniques are subsequently described, with a concentration on LSA techniques.

Related literature

Summary writing

Summarization involves two major cognitive processes. One process consists of *comprehension* of the source material, which is related to deeply understanding the subject matter. More specifically, this process and the associated abilities consist of identifying main ideas, distinguishing them from supporting details, and identifying rhetorical structures and the organization of original source texts (E. Kintsch, 1990; Spirgel & Delaney, 2016). According to W. Kintsch's (1998) construction–integration model, this requires (1) the construction of local and global relationships among the propositions that are extracted from the explicit textbase, and (2) the construction of a situation model of the text meaning, based on both the textbase and relevant prior knowledge, to bridge conceptual gaps with inferences (van Dijk & Kintsch, 1983).

The other major process is *content reproduction*, which involves generalization, synthesis, and coherently organized writing. In particular, this process and the associated abilities include selecting the most important information, deleting unimportant details and redundant information, substituting more abstract superordinate expressions, selecting or inventing a topic sentence (Brown & Day, 1983; van Dijk & Kintsch, 1983), and succinctly communicating this information in writing (León, Olmos, Escudero, Cañas, & Salmerón, 2006).

Empirical evidence has demonstrated that summarization is an effective strategy to foster deep reading comprehension and learning (W. Kintsch, 1998; Spirgel & Delaney, 2016; van Dijk & Kintsch, 1983). For example, summary writing helps students encode and strengthen the retention of new information (Hinze & Rapp, 2014; Karpicke & Roediger, 2007; Stewart, Myers, & Culley, 2010). With respect to noncognitive benefits, summary writing can reduce students' test anxiety, as compared to traditional testing (Mok & Chan, 2016).

Empirical evidence supports the conclusion that summarization strategies are effective for different types of learners, including native speakers (Britt & Sommer, 2004; Leopold, Sumfleth, & Leutner, 2013; Trabasso & Bouchard, 2002; Westby, Culatta, Lawrence, & Hall-Kenyon, 2010), language learners (Baleghizadeh & Babapur, 2011; Chiu, 2015; Oded & Walters, 2001; Shokrpour, Sadeghi, & Seddigh, 2013), students with learning disabilities (Jitendra, Cole, Hoppes, & Wilson, 1998; Jitendra, Hoppes, & Xin, 2000; Rogevich & Perin, 2008), and students with low literacy skills (Perin & Lauterbach, 2016; Perin, Lauterbach, Raufman, & Kalamkarian, 2016). Moreover, summarization is more efficient and effective in improving student learning than are other formats of assessment, such as short-answer comprehension questions (Carroll, 2008; Shokrpour et al., 2013), argument essay writing (Gil, Bråten, Vidal-Abarca, & Strømsø, 2010), multiple-choice questions, and fill-in-the-blank questions (Mok & Chan, 2016).

Even though summarization is an effective instructional strategy and a good measure of deep learning, school teachers often do not adopt summary writing as a primary means of assessment, because manually grading summaries is enormously time-consuming. Fortunately, researchers from the fields of machine learning and natural language processing (NLP) have developed approaches to computerized summary scoring (CSS) to measure the quality of the content and style of summaries. The next three sections review some tools that are used in CSS, and each section is named according to the prominent features of the techniques, such as machine learning, NLP, and LSA, but with a focus on LSA. This means that, for instance, machine-learning techniques may be involved in techniques we classify under NLP, but the unique characteristics of the technique derive from NLP.

Computerized summary scoring: machine learning

Researchers in previous studies on CSS have used some techniques from machine language translation and information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). One primary technique was *n*-gram co-occurrence (Lin & Hovy, 2003), which was adapted from BLEU (bilingual evaluation understudy; Papineni, Roukos, Ward, & Zhu, 2002), a technique for machine translation. Later, researchers revised and modified this method and developed ROUGE (recall-oriented understudy for gisting evaluation; Lin, 2004) for summary assessment. Both of these tools compared a target summary with one model summary or multiple summaries written by humans. Both methods were evaluated by counting the percentage of *n*-gram co-occurrences. They found that multiple model summaries led to more stable evaluation results. Pearson correlations between automated methods and human scores for the single document task were .35 on average, ranging from .02 to .91 with BLEU techniques to grade 100–word summaries, but improved with ROUGE techniques.

To improve the accuracy and performance, a pyramid method compared the target summary with a pool of human summaries rather than with one model summary (Nenkova & Passonneau, 2004). These model summaries were manually annotated by human experts on the basis of summarization content units (SCUs). Experts first located similar sentences and then identified more closely related subparts, using a clause as the maximal linguistic unit. Differential weights were assigned to semantic matching content units based on their frequency in a corpus of summaries. As the size of the corpus increases, the number of SCUs grows and it led to more stable and meaningful content evaluation. The pyramid method improved the CSS performance when compared to BLEU and ROUGE, but it required substantial human effort for annotation. In the long run, it is impractical for school teachers to use these CSS tools. Not only does annotation take much more time than manually grading, but also model validation requires the professional skills for machine learning techniques.

Computerized summary scoring: natural language processing

Coh-Metrix is a popular tool that extracts more than a hundred linguistic and discourse measures by adopting advances in computational linguistics, NLP, and machine learning (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014). Coh-Metrex was originally designed to analyze the complexity of edited texts and conversations. Recently, some researchers have also used Coh-Metrix measures to predict human summary scores. For example, Perin and Lauterbach (2016) used the measures that fit for essay assessment in the study of McNamara, Crossley, and McCarthy (2010) as well as a larger set of 52 Coh-Metrix measures. When using the full set of Coh-Metrix measures, they found that three Coh-Metrix measures predicted 22% of the variance in human analytic scores on written summaries with a stepwise regression: referential cohesion (content word overlap), lexical diversity (richness of vocabulary usage), and word information (familiarity).

Li et al. (2016) conducted a similar study but added more Coh-Metrix measures, for a total of 94 measures. It is beyond the scope of the present study to describe these measures and the major results, but some results are particularly relevant to the present study. Li, Cai, and Graesser reported that 55 Coh-Metrix measures significantly predicted 18% of the variance in human scores; these measures related to referential cohesion (e.g., noun/argument/stem overlaps), LSA overlap (e.g., adjacent sentences, given/new), lexical diversity (e.g., type–token ratio), connectives (e.g., logical, additive), situation model (e.g., causal verbs and particles, LSA verb overlap), syntactic complexity (e.g., minimal edit distance, sentence syntactic similarity), syntactic pattern density (e.g., noun/verb/preposition phrase density), word information (e.g., noun, adjective, hypernym for nouns), and readability measured by the Flesch–Kincaid grade gevel (FKGL; Klare, 1974–1975).

These two studies suggest that there is some fluctuation in which Coh-Metrix measures predict human summary scores, even though the amounts of variance explained were comparable. The measures apparently vary somewhat with the change of original source texts. Consequently, model training and model validation are needed when source texts are changed. Similar to machine learning tools, the Coh-Metrix approach to CSS is also impractical and time-demanding for school teachers because it requires complex professional understanding of NLP, model training, and model validation.

Computerized summary scoring: latent semantic analysis

LSA is a mathematical method for representing the meaning of words and text segments based on a large corpus of text (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). A large two-dimensional matrix is generated, in which each row represents a word (term) and each column represents one text segment (document). LSA uses singular value decomposition (SVD), a mathematical approach in linear algebra, to reduce the matrix to a much smaller set of *K*-dimensions (typically, 100 to 500; Landauer & Dumais, 1997). Each word and each document are represented by a vector with *K*-dimensions in the semantic space.

The extent to which the two documents are semantically associated can be determined by comparing the semantic information contained in the document units or in entire documents. The units may be words, phrases, clauses, sentences, paragraphs (Foltz, 1996; Landauer, 1998; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), or summaries (Foltz, 1996; E. Kintsch et al., 2000; Li et al., 2016; Olmos, León, Escudero, & Jorge-Botana, 2011; Olmos, León, Jorge-Botana, & Escudero, 2009, 2013). LSA uses the geometric cosine between two vectors to compute the conceptual similarity between any two units of the text. Meaningful cosine values range from approximately 0 (representing *no similarity*) to 1 (representing *perfect similarity*).

Researchers have frequently used LSA to automatically evaluate students' written summaries. Empirical evidence has shown that LSA robustly predicts the semantic quality of summaries as perceived by humans (Jorge-Botana, Luzón, Gómez-Veiga, & Martín-Cordero, 2015; E. Kintsch et al., 2000; Li et al., 2016; Olmos et al., 2011, 2013; Sung, Liao, Chang, Chen, & Chang, 2016; Wade-Stein & Kintsch, 2004). One successful system that uses LSA to automatically assess summaries is Summary Street® (E. Kintsch et al., 2000; Wade-Stein & Kintsch, 2004). Summary Street uses LSA to provide students feedback on written summaries in terms of content, redundancy, and relevance through a graphic

representation. This system allows students to modify and resubmit summaries. Evidence showed that Summary Street substantially improved the quality of students' summaries, especially when the subject matter had a higher difficulty level (E. Kintsch et al., 2000). This method has been successfully applied to other systems, such as WriteToLearn (Landauer, Lochbaum, & Dooley, 2009), distance education (Jorge-Botana et al., 2015), and an online summary-assessment-and-feedback system (Sung et al., 2016).

Some previous studies have found that the satisfactory performance of LSA was restricted to longer essays, such as essays more than 60 words (Rehder et al., 1998; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999), and to expository rather than narrative texts (León et al., 2006). To address these limitations, some researchers developed new LSA algorithms in order to improve the reliability of LSA to capture the meaning of short summaries (25–50 words). Olmos et al. (2009) and Olmos et al. (2011) compared three new algorithms: a semantic common network algorithm (W. Kintsch, 2001, 2002), a best-dimension reduction measure of the latent semantic space (Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007), and the Euclidean distance measure (Rehder et al., 1998). The results showed that a best-dimension algorithm best predicted the short expository summaries, as compared to other LSA algorithms.

Researchers have also compared two other LSA methods that fall into two major categories: holistic and analytic (also called *componential*). The holistic method provides a summary score based on the overall similarity to the model summary, whereas the analytic method computes similarity scores based on sublevel components of the summary (e.g., individual sentences, coherence, content, main topic) as compared to the model summary. The holistic method showed higher accuracy for the overall quality of a summary, whereas the analytic method had the advantage of providing more specific details in terms of the sublevel components (Foltz, Gilliam, & Kendall, 2000). Previous research compared five holistic methods (crowdsourcing summary, individual summaries, source text, expert summary, pregraded summaries) and two analytic methods (sentence vs. main sentence) (Li et al., 2016; Olmos et al., 2011). The analytic method either calculated the cosine scores between a target summary and each sentence in the original source text or computed the cosine scores between each sentence in a target summary and a set of important sentences that experts had identified in the source text. The present study explores LSA methods to predict overall summary quality, so only the holistic methods are addressed, as we describe below.

Crowdsourcing summary Crowdsourcing enables a diverse and potentially large population (here, students or learners) to generate a large set of summaries (Li et al., 2016). The crowdsourcing holistic LSA method computes the LSA cosine score to predict summary scores. This method compares the target summary with a crowdsourcing model summary, which includes a set of all other summaries that need to be graded. The target summary is a single summary to be graded. The crowdsourcing model summary is a very large document in which all the other summaries are concatenated. Suppose that 100 summaries are to be graded and Tom's summary is the first one to be graded. His summary is selected and treated as the target summary, whereas the remaining 99 summaries written by other students are concatenated into one document, called the crowdsourcing model summary. If Mary's summary is the second one to be graded, her summary is taken out from the crowdsourcing model summary and treated as the target summary. Meanwhile, Tom's summary is no longer the target summary, but is concatenated into the crowdsourcing model summary. In this way, the crowdsourcing model summary is always the concatenation of 99 summaries.

Li et al. (2016) compared the crowdsourcing LSA method with NLP methods, such as Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014) measures, LIWC (Pennebaker, Boyd, Jordan, & Blackburn, 2015) measures, and a combination of Coh-Metrix and LIWC measures. The results showed that the crowdsourcing LSA had predictability for human summary scores equivalent to those of the other three NLP methods, with moderate correlations (r = .44) on average. Empirical evidence will be needed to confirm the extent to which crowdsourcing LSA (good, medium, poor, or all crowdsourcing summaries) predicts human summary scores as compared to other LSA methods (good, medium, and poor expert summaries along with source texts), which was the goal of this study.

Individual summary Olmos et al. (2011) proposed another alternative LSA method that used students' summaries. This method compared the target summary with each individual student's summary and then the average cosine score was assigned to the target summary. To differentiate this method from the crowdsourcing summary, we call this the individual summary method. In essence, this method is similar to the crowdsourcing summary that we proposed, except that the comparison is made with each individual summary separately. Olmos et al. (2011) found that this method did not robustly predict human summary scores relative to an expert summary and pregraded summaries.

Source text The source-text-based holistic method calculates the LSA cosine by comparing each target summary with the entire original source text. Summary Street applied this method in order to assess written summaries of longer source texts with about a thousand words (E. Kintsch et al., 2000; Wade-Stein & Kintsch, 2004). This method involved three steps. First, the source text was divided into different sections based

on subheadings or subtopics. Second, in each section, the LSA cosine scores of each sentence with other sentences were computed and the highest LSA cosine score in the section was identified; the typical sentence represents the content of the section. Third, the typical sentences in all the sections were formed into a typical summary, which was used as a model summary and represented the meaning of the source text. The performance of this method, based on correlations with human scores, varied from .32 to .88, depending on topics and genres of the source texts. Olmos et al. (2011) directly calculated the LSA cosine scores between a target summary and the source text and found that this method was less robust in predicting human summary scores than were expert summary and pregraded summaries.

Expert summary The expert-summary-based holistic method compares the target summary with a single expert summary or with several summaries written by different experts. The reason for using several expert summaries rather than just one as the model is that different experts may construct summaries with different content items that they assume to be important and succinct. The more expert summaries that are generated as the model summary, the more the prediction of human summary scores may improve. Typically, researchers have four or five experts write the ideal summaries and then assign a best fit score, namely the highest LSA cosine score between the target summary and the expert summary. Olmos et al. (2011) reported that expert summaries ($n = 4$) predicted human scores better than the source text and than individual summaries, and the method was as good as the pregraded summaries. However, in the long run, the expert LSA method is impractical for school teachers or small-scale assessment, due to the added effort of collecting summaries from experts. If teachers use summary writing as a primary, weekly assessment for deep reading, they would need to find other experts to help them generate expert summaries each week, and there is typically no budget to hire these experts.

Pregraded summary The pregraded-summary-based holistic method was proposed by E. Kintsch et al. (2000) when they developed Summary Street. This method calculates the LSA cosine by comparing the target summary with a set of pregraded summaries. Human effort is needed to manually score these summaries, so this method was not applied to Summary Street. Olmos et al. (2011) generated a sample of summaries that were first graded by 100 experts and then compared the target summary with each pregraded summary. The target summary was assigned the average score of a set of ten closely similar summaries, weighted by their similarity score. They found that the pregraded summaries could predict human summary scores as well as expert summaries, both of which had a better predictability than the source text and individual summaries. Their studies showed similar predictability to studies that were conducted by E. Kintsch et al. (2000), with the correlation ranging from .41 to .63 for the expository text and .46 to .58 for the narrative text. However, this approach requires some experts to pregrade 100 summaries, which makes it impractical for school teachers to assess summaries in classes, large or small. For small classes, of less than 100 students, there are not enough summaries for modeling. For large classes, the teacher needs to grade 100 summaries, which is obviously labor intensive. It is not clear whether this method works for fewer number of summaries in modeling.

Among these five holistic LSA methods, only the first three (i.e., crowdsourcing summary, individual summary, and source text) do not require teacher effort to generate the model summaries or to grade summaries. Moreover, the latter two methods have been shown to be less robust than the expert-summary and pregraded-summary methods (Olmos et al., 2011). Expert-summary LSA is considered to be the best method, because it predicts human scores better than the source text and individual-summary methods and requires less human effort than the pregraded-summary method. However, it is unresolved whether the crowdsourcing method predicts human scores as well as the expert-summary method. Therefore, empirical evidence is needed before we can conclude that the expert summary method is the best LSA method for CSS.

Research questions

The present article aims to answer three research questions:

1. Does crowdsourcing LSA predict the content of summaries (as perceived by humans) as well as, if not better than the expert summary method and the source text method?
2. Does the number of crowdsourcing summaries in the crowdsourcing LSA method affect the prediction of human summary scores?
3. Is there an optimal number of crowdsourcing summaries for the crowdsourcing LSA method in predicting human summary scores?

To answer the first question, we compared crowdsourcing LSA with expert-summary LSA and the source text. We did not include the individual-summary method because this method has been shown to be less predictive than the expert summary method (Olmos et al., 2011). We did not include pregraded summaries because this method is time-consuming and impractical to implement in educational contexts. However, we did include another three crowdsourcing LSA methods, in order to consider three levels of summary quality: good, intermediate, or poor. These summaries were indeed pregraded, but they were added in order to evaluate whether the quality of crowdsourcing summaries matters. If

quality does matter, the pregraded crowdsourcing LSA would be a more predictive method, even though it involves added costs and effort as compared to the expert summary LSA. If quality does not matter, all of the crowdsourcing summaries except the target summary could be concatenated together as one model summary without pregrading, and thereby save considerable human effort and costs.

We include the source-text method for two reasons. First, we assume that this method can predict human summary scores to some extent, but the level of predictability is unclear, because it was not reported in the study by Olmos et al. (2011). Second, Olmos et al. (2011) did not state the length of the source text, but only mentioned that the length of summaries ranged from 25 to 50 words. Therefore, it is necessary to include the source-text method as a comparison to understand the explicit role of the source text method in CSS. We predicted that crowdsourcing LSA could predict human summary scores as well as, if not better than, the expert-summary method and source-text method, because the crowdsourcing summary contains rich content that captures more information presented in the diverse students' summaries than does one expert summary or one source text.

If crowdsourcing LSA is approved as a valid method for CSS, the next questions that researchers and practitioners will be concerned with are whether the number of crowdsourcing summaries affects performance and whether a certain number yields optimal performance. To answer these questions, we compared the performance of crowdsourcing LSA with different numbers of summaries in the crowdsourcing model summary. We hypothesized that the number of summaries might not affect the performance of the crowdsourcing method when the number of summaries is sufficient. When any sufficient number of summaries are included, the content in the target summary may be captured by the crowdsourcing model summary. Adding more summaries to the crowdsourcing model summary, over and above the minimal number, would increment the content only a negligible amount. Consequently, increasing the number of summaries might not improve the performance after the minimal number needed had been reached.

This study makes two contributions to the literature. First, we provide an efficient and effective approach to CSS, because crowdsourcing LSA does not require any human effort for the generation of a model summary. The only requirement is to have a particular minimal number of written summaries to be graded, which is always the case in courses in which students are expected to turn in summaries. This method renders it unnecessary for experts to either manually pregrade summaries or write model summaries. Instead, the same summaries that need to be scored can be used as the model summary, which saves time, human effort, and costs. Second, this study compares three methods: Expert summary is considered the most valid method, source text is inferior to expert summary, and it is unclear how well crowdsourcing LSA will predict human summary scores. This study compares the performance of these three methods of CSS in predicting human summary ratings, while also considering the quality of the summaries.

Method

Participants

Crowd workers ($N = 240$) volunteered for monetary compensation ($30 for writing eight summaries) on Amazon Mechanical Turk (AMT), a trusted and commonly used data collection service (Paolacci & Chandler, 2014). High-quality data are obtained inexpensively and rapidly through AMT, with more diverse demographics than standard Internet samples and typical American college samples (Buhrmester, Kwang, & Gosling, 2011). For example, self-reports of individual differences (i.e., age and gender) are psychometrically valid (Buhrmester et al., 2011; Shapiro, Chandler, & Mueller, 2013). The quality of linguistic judgments that Turkers provide is comparable to that of college samples (Sprouse, 2011). The data obtained through AMT are at least as reliable as those obtained via traditional methods according to Buhrmester et al. (2011).

The basic requirement for participation is that the participants have the desire to improve English summary writing. Due to the technical issues and other unidentified reasons, 21 participants could not complete the experiments. Human graders discovered that 18 participants copied the content from Wikipedia so their corresponding 47 summaries were removed. Thus, 201 workers submitted 1,480 summaries for eight expository texts. The average age of the participants was 33.50 with a standard deviation of 8.79 (57% male; 81% with a bachelor's degree or above; 89% non-English speakers with at least 18 years of English learning). Of the participants, 71% were Asian, 16% white or Caucasian, 7% African American, 5% Hispanic, 2% other.

Materials

Participants read eight short expository texts (195 to 399 words) with different topics and text difficulties in AutoTutor CSAL, an intelligent tutoring system (ITS) to enhance adults' reading comprehension by teaching summarization strategies (Li et al., 2016; Li, Cheng, Yu, & Graesser, 2015; Li & Graesser, 2017; Li, Shubeck, & Graesser, 2016). After reading each text, participants were asked to write a summary with 50–100 words. Four texts had a compare–contrast text structure and another four had a cause–effect text structure. The text difficulty was measured by Coh-Metrix

formality (Graesser et al., 2014; Li, Graesser, & Cai, 2013) and Flesch–Kincaid grade level (FKGL; Klare, 1974–1975). Coh-Metrix formality measures language style that ranges from informal discourse to formal discourse at multiple textual levels with five primary Coh-Metrix components (Graesser et al., 2014; McNamara et al., 2014). Formality increases with word abstractness, syntactic complexity, expository texts, and high referential cohesion and deep cohesion. Coh-Metrix formality has a standardized score with the TASA corpus as a reference corpus (Graesser et al., 2004, 2014; McNamara et al., 2014), with numbers higher than 0 representing more formal discourse, and numbers below 0 representing more informal discourse. This measure more comprehensively reflects text difficulty based on words, syntax, cohesion, and genres, as compared to FKGL, which focuses on shallow textual features, such as sentence length and word length (Graesser et al., 2004). The eight expository texts tended to be more formal, with formality ranging from .12 to .64. On the basis of the FKGL, these texts are suitable for students between grades 8 and 12. Table 1 displays the text difficulty scores of the source texts and the participants' written summaries, as measured by formality, FKGL, and word count. The word count column for the summaries displays the average word count of all summaries for each source text.

Rubric for summary grading

The rubric for summary grading was constructed on the basis of the rubrics proposed by Garner and McCaleb (1985) and Friend (2001), with a slight modification described in Li et al. (2016) and Li and Graesser (2017). Two components in previous rubrics remained intact: (1) summarization thesis statement and (2) content inclusion and exclusion. We included a basic component that is frequently used when grading summaries and essays by humans: grammar and mechanics. We also added a component, signal words for text structures, because participants were trained in this study to use signal words to express comparisons (e.g., *similarly*, *however*) and cause–effect relationships (e.g., *consequently*, *therefore*). Conversely, we removed the sentence transformation that was adopted in previous rubrics because the system automatically detected copying if students copied ten consecutive words from the source text in their summaries. If this occurred, the system did not allow the summary to be submitted. The maximum score for each component was 2 points, and the minimum was 0 points. Thus, the total score ranged from a minimum of 0 to a maximum of 8 (see Table 2).

Human grading

Four English native senior Ph.D. students graded summaries, one male and three females. Before grading, they were given 10 min to discuss and get familiar with the rubric displayed in Table 2. Then, three rounds of training for summary grading were conducted, with one round per week. For each round, we randomly selected four summaries from each source text with a total of 32 summaries from eight texts. Four graders scored 32 summaries independently. After each round of grading, they discussed discrepancies before going to the next round. Interrater reliability was assessed by the intraclass correlation coefficient with a two-way random model and absolute agreement type (Shrout & Fleiss, 1979). The average interrater reliability reached the threshold: Cronbach's α = .82, intraclass correlation coefficient = .80. The average of the reliabilities for the three training rounds was high, so the four graders scored the remaining summaries. Specifically, each grader scored the summaries from two source texts of the same text structure. These human summary scores were used as the gold standard for evaluation of the LSA methods.

LSA corpus

In this study, the LSA corpus contained 37,520 texts created by TASA (Touchstone Applied Science Associates, Inc.,

Table 1 Text characteristics of the original source texts and the written summaries

Text structure	Topics	Source text			Summaries (N = 1,480)			
		Formality	FKGL	W.C.	Formality	FKGL	W.C.	N
Comparison	Butterfly and Moth	.12	8.6	255	0.07	9.6	71.9	183
	Hurricane	.20	9.4	222	0.44	12.0	71.3	185
	Walking and Running	.18	8.9	399	0.15	10.2	71.3	187
	Kobe and Jordan	.14	9.2	299	0.09	8.8	68.3	187
Causation	Floods	.47	9.2	230	0.45	9.1	69.1	186
	Job Market	.62	10.9	240	0.44	11.6	68.1	181
	Effects of Exercising	.28	9.1	195	0.15	10.5	62.8	189
	Diabetes	.64	11.7	241	0.49	11.0	73.7	182

W.C. = word count. FKGL = Flesch–Kincaid grade level

Table 2 Rubric for summary grading

Components	0 points	1 point	2 points
Thesis statement	The summary does not state the main ideas.	A topic sentence that touches upon the main ideas.	A clear topic sentence that states the main ideas.
Content inclusion & exclusion	Few pieces of major supporting information stated and not necessarily in a logical order. Many minor or unimportant details or reflections.	Some but not all major supporting information stated and not necessarily in a logical order. Some minor or unimportant details or reflections.	Major supporting information stated economically and arranged in a logical order. No minor or unimportant details or reflections.
Mechanics & grammar	Serious errors in mechanics, usage, grammar or spelling, which make the summary difficult to understand.	Some errors in mechanics, usage, grammar or spelling that to some extent interfere with meaning.	Few or no errors in mechanics, usage, grammar or spelling.
Signal words	Uses several clear signal words to connect information.	Uses several clear and accurate signal words to connect information.	Uses the clear and accurate signal words to connect information.

renamed Questar Assessment Inc.). Texts in the TASA corpus have nine genres, with Language Arts, Science, and Social Studies as the three major genres, which span across 13 grade levels, from kindergarten through 12th grade. Texts have a mean length of 289.0 words (SD = 24.6), with a total number of 10,880,726 words. The number of dimensions was set at 300.

Procedures

Crowdsourcing summary We compared crowdsourcing LSA with seven other LSA holistic methods: source text, expert-generated summaries (of good, intermediate, and poor quality), and crowdsourcing summaries (of good, intermediate, and poor quality). The crowdsourcing summary was generated by including all the summaries that were written by participants except the summary that was to be graded. We call this type of crowdsourcing summary *Crowdsourcing-All*. People may have concerns that mixing the good, intermediate, and poor summaries may increase unwanted noise and reduce predictability. Therefore, we coded students' summaries into three levels of quality to investigate whether this is a problematic issue. Specifically, all the crowdsourcing summaries of good quality, with human scores from 6 to 8 points, were put into one document, which was called *Crowdsourcing-Good*. Another document contained all the intermediate crowdsourcing summaries, with scores from 3 to 5 points, and was called *Crowdsourcing-Intermediate*. The third document consisted of all of the poor crowdsourcing summaries, from 0 to 2 points, and was called *Crowdsourcing-Bad*. The summary content scores of the three levels of quality were significantly different, $F(2, 1437) = 554.70$, $p < .001$, with the means of 0.31 (SD = 0.46, N = 165), 1.00 (SD = 0.48, N = 751), and 1.63 (SD = 0.48, N = 524) for poor, intermediate, and good, respectively. In total, these four types of crowdsourcing model summaries were compared with expert-generated summaries and the source text. The text difficulty of the Crowdsourcing-Good, -Intermediate, and -Poor documents are listed in Table 3. A small number of unusual summaries were removed from the observations that were analyzed. We removed five summaries with gibberish content—for instance, if a participant used the same letter consecutively (such as "zzzzzzz").

To answer the first question, we randomly selected 180 summaries from each source text, a maximal number set to maintain the same number of summaries for each source text. Overall, 533 summaries (36%) were good, 772 (52%) were intermediate, and 175 (12%) were poor. Their text difficulty scores were similar across all summaries, as is displayed in Table 1.

To answer the second and third questions, of whether the number of summaries in the crowdsourcing model summary affects the performance of crowdsourcing LSA and what number yields optimal performance, we generated two subsets of summaries: a model summary subset and a target summary subset. Specifically, we randomly selected 100 summaries in order to generate a model summary subset. Thus, the remaining 80 summaries were put into a target summary subset. The summaries in the model subset were used to generate the crowdsourcing summary, whereas the summaries in the target subset were to be scored to evaluate the performance of the crowdsourcing LSA methods. The model subset and the target subset had similar proportions of summary qualities (good, intermediate, and poor) relative to the overall summaries. We do not list the text difficulty scores in Table 4, because they were similar across the levels of summary quality (see Table 3) and also across all summaries (see Table 1). Therefore, Table 4 only displays the numbers of summaries in each subset in terms of the three levels of summary quality.

Furthermore, we randomly selected ten summaries from the 100-summary model set (n = 10) and labeled that subset as "10" (n = 10). Then we randomly selected another ten summaries, added them to the previous Subset "10," and labeled the expanded subset as "20" (n = 20). This

Table 3 Text characteristics of the crowdsourcing summaries (N = 1,440; N = 180 for each topic): Good, intermediate, and poor

Text structure	Topics	Good				Intermediate				Poor			
		Formality	FKGL	W.C.	N	Formality	FKGL	W.C.	N	Formality	FKGL	W.C.	N
Comparison (N = 742)	Butterfly and Moth	0.09	9.1	76.1	81	0.05	9.7	70.0	89	0.12	11.4	58.5	13
	Hurricane	0.67	11.9	72.8	54	0.51	12.5	73.1	79	0.10	11.6	66.8	52
	Walking and Running	0.15	10.3	75.9	83	0.14	10.4	68.2	98	0.18	7.5	57.0	6
	Kobe and Jordan	0.22	9.3	75.8	50	0.06	8.7	66.7	110	– 0.03	8.4	60.7	27
Causation (N = 738)	Floods	0.53	8.8	72.4	47	0.45	9.3	69.1	112	0.28	8.9	63.6	27
	Job Market	0.40	11.2	69.3	53	0.49	11.6	70.1	101	0.30	12.1	58.0	27
	Effects of Exercising	0.10	9.9	66.9	86	0.22	11.4	61.1	92	0.03	8.8	45.2	11
	Diabetes	0.55	11.4	77.4	79	0.47	10.1	72.3	91	0.30	16.4	60.4	12

W.C. = Word Count.

continued until all the summaries had been randomly selected. Thus, ten subsets were generated, each labeled as 10, 20, . . . , 100. With these ten subsets, we investigated whether changing the number of crowdsourcing summaries affected the predication of the human summary scores. We investigated the effect of increasing the number of crowdsourcing summaries from 10 to 100 in order to simulate the increases of teachers' students from year to year. For example, if a teacher had a class of 20 students this year, 40 students next year, and so on, the teacher could cumulatively use the summaries to get better models.

Source text and expert summary The source text was directly used as the model summary. As for the expert summary, one doctoral student wrote three summaries with good, intermediate, and poor qualities for each source text based on the rubric. Another doctoral student revised and modified these summaries. The crowdsourcing summaries were classified into three levels of quality, so we also prepared three levels of quality of the expert summaries, called Expert-Good, Expert-Intermediate, and Expert-Poor. This allowed us to examine whether the three quality levels of the summaries written by experts would show results similar to the three quality levels of the crowdsourcing summaries.

In total, this study compared eight LSA methods as follows:

(1) Expert-Good: ideal summary written by experts;
(2) Expert-Intermediate: summary written by experts to have intermediate quality;
(3) Expert-Bad: summary written by experts to have poor quality;
(4) Source Text: original text that participants read to write summaries;
(5) Crowdsourcing-Good: students' summaries that were scored as good by experts;
(6) Crowdsourcing-Intermediate: students' summaries that were evaluated as intermediate;
(7) Crowdsourcing-Bad: students' summaries that were graded as poor; and
(8) Crowdsourcing-All: all students' summaries without considering summary quality.

Table 4 Number of summaries: Model summary corpus, and target summary corpus

Structure	Topics	Total (n = 180)			Model (n = 100)			Target (n = 80)		
		Bad	Inter.	Good	Bad	Inter.	Good	Bad	Inter.	Good
Comparison	Butterfly and Moth	12	88	80	7	49	44	5	39	36
	Hurricane	50	78	52	28	43	29	22	35	23
	Walking and Running	5	95	80	3	53	44	2	42	36
	Kobe and Jordan	25	105	50	14	58	28	11	47	22
Causation	Flood	26	108	46	14	60	26	12	48	20
	Job Market	26	101	53	14	57	29	12	44	24
	Effects of Exercising	10	86	84	5	48	47	5	38	37
	Diabetes	11	90	79	6	50	44	5	40	35

Inter. = Intermediate.

Results

The average summary score was 4.49 (*SD* = 1.55) out of 8. The average writing time was 416.8 s (*SD* = 9.8) for each summary. Pearson correlations have been frequently used in previous studies to evaluate the accuracy of automated scoring (Li et al., 2016; Lin & Hovy, 2003; Olmos et al., 2011). In this study, we used Pearson correlations to compare the performance between each of the eight LSA methods and human summary scores. Fisher's *z* was used to compare whether one correlation was statistically different from another. We also compared the minimum and maximum correlations across the eight topics because they were likely to reflect the variation in performance across the topics for each method. Specifically, a small range with higher correlation scores would indicate better method performance. All the LSA cosine scores were standardized on the basis of each source text, with a mean of 0 and standard deviation of 1. The reason why we used the standardized scores was to compare LSA cosine scores across different source texts and to have equal weighting of different texts.

We first examined how well crowdsourcing LSA predicted the content of the summaries, as compared to the expert summaries and source text. Then we investigated whether the number of crowdsourcing summaries affected the performance and what number yielded optimal performance.

Predictability of crowdsourcing LSA

Table 5 displays the Pearson correlations between the LSA standardized cosine scores (hereafter called *LSA score*s) of the eight CSS methods and the human summary scores. All correlations were statistically significant, with *r*s ranging from .24 to .51. On the basis of the average of the correlations across the eight topics, the order of performance from high to low was: (1) Crowdsourcing-Good, (2) Expert-Good, (3) Crowdsourcing-All, (4) Crowdsourcing-Intermediate, (5) Source Text, (6) Crowdsourcing-Poor, (7) Expert-Intermediate, and (8) Expert-Poor. The top two methods required human effort for either pregrading or generating good expert summaries. Crowdsourcing-All, requiring no human effort to create the model, was in third place.

We performed Fisher's *z* analyses to see whether the correlations were significantly different. The results indicated that the correlations between the Crowdsourcing-All LSA cosine scores and human scores were not significantly different from the correlations between human scores and the other LSA methods that involved human scorers (Crowdsourcing-Good and Expert-Good), but was significantly higher than Expert-Intermediate for the topic of *Diabetes*, $z = 2.14$, $p < .05$, and than Expert-Poor for two topics: *Walking and Running*, $z = 2.57$, $p < .01$, and *Effect of Exercise*, $z = 2.14$, $p < .05$. The Fisher's *z* results suggested that the Crowdsourcing-All LSA method could predict human summary scores as well as Crowdsourcing-Good and Expert-Good, but better than Expert-Intermediate and Expert-Poor.

We found no significant differences in the correlations among the four crowdsourcing methods, source text, and Expert-Good summary; however, comparisons of the minimum and maximum correlations revealed that Crowdsourcing-All (.40–.49), Crowdsourcing-Good (.41–.51), and Expert-Good (.39–.51) all had better performance than Crowdsourcing-Intermediate (.37–.50), Crowdsourcing-Poor (.32–.43), and source text (.29–.48) across the eight topics (see Table 5). The latter three methods had low correlations of .29–.37, whereas the former three methods were all at least .39.

Moreover, Crowdsourcing-All LSA was almost perfectly correlated with Crowdsourcing-Good LSA, with *r*s ranging from .97 to 1.00 across the eight topics, and was also highly correlated with Expert-Good LSA, with an average score above .80 (*r*s = .59–.95). Therefore, Crowdsourcing-All LSA is an ideal method for CSS because it does not require experts to pregrade students' summaries or write gold-standard summaries, both of which add significant amounts

Table 5 Correlations between human summary scores and eight LSA cosine scores ($N = 1{,}440$)

Structure	Topic	C_A	C_G	C_I	C_P	Text	E_G	E_I	E_P
Comparison	Butterfly and Moth	.40	.41	.37	.32	.29	.40	.31	0.24
	Hurricane	.43	.49	.44	.39	.45	.47	.41	0.45
	Walking and Running	.48	.50	.46	.34	.41	.39	.42	0.25
	Kobe and Jordan	.45	.48	.46	.39	.34	.49	.43	0.43
Causation	Floods	.43	.49	.45	.37	.44	.45	.36	0.42
	Job Market	.40	.44	.40	.33	.38	.41	.33	0.29
	Effect of Exercising	.46	.48	.44	.37	.43	.44	.36	0.27
	Diabetes	.49	.51	.50	.43	.48	.51	.30	0.38
Average		.44	.48	.44	.37	.40	.45	.37	.34

All the correlations were significant, *p*s < .01. C = Crowdsourcing; A = All; G = Good; I = Intermediate; P = Poor; E = Expert; Text = Source Text. $N = 180$ for each text

of human effort, time, and cost. The answer to the first question is that crowdsourcing LSA predicted the content of summaries as well as Crowdsourcing-Good, Crowdsourcing-Intermediate, and the expert summary, as well as moderately better than the source-text, Crowdsourcing-Poor, Expert-Intermediate, and Expert-Poor methods.

Test of sample size

The results reported above were based on 180 summaries for each source text. This large number of summaries is unlikely to be obtained by a school teacher, especially at a small school. If a school teacher has one or two classes, the teacher might have only about 20–40 summaries. This section compares the performance of crowdsourcing LSA with different numbers of crowdsourcing summaries and examines whether the number of the crowdsourcing summaries affects performance, with sample size increasing in increments of ten (sample size starts from ten summaries).

The results showed that correlations were good and that the majority of correlations were almost perfect (rs = .88–1.00) between pairs of crowdsourcing LSA cosine scores that were computed with different numbers of crowdsourcing summaries (N = 10, 20, . . . , 100) across the eight topics (see Table 6). This finding implies that the crowdsourcing LSA scores with a small number of crowdsourcing summaries (e.g., ten) were similar to the scores with a large number of crowdsourcing summaries (e.g., 100).

Table 7 presents the correlations between the crowdsourcing summaries and human summary scores across the eight topics as a function of the number of summaries in the crowdsourcing summary, ranging from 10 to 100 at intervals of ten summaries. The averages of the correlations between crowdsourcing LSA and human scores across the eight topics were medium, ranging from .37 to .43 as the number of crowdsourcing summaries went from 10 to 100. As Table 7 illustrates, the correlations tended to gradually increase as summaries increased, even though these correlations were not significantly different. The results also showed consistent correlations within a topic, even though the numbers of crowdsourcing summaries were different.

All these findings indicated that the number of crowdsourcing summaries did not appreciably matter. The performance of ten crowdsourcing summaries was similar to that with more summaries, up to 100. However, when more summaries were included, we observed a slight increase in performance, especially when the sample size reached a level of at least 40.

Discussion

For the present study we developed and evaluated a crowdsourcing-based LSA approach to automatically score summaries without any human involvement through either grading or generating summaries to produce a model summary. The model summary was instead generated by a crowd population, so we called this method the *crowdsourcing LSA* method. This method was evaluated by comparing it with seven other LSA methods, which consisted of using the source text, expert-generated summaries (with good, intermediate, and poor quality), and crowdsourcing summaries (with good, intermediate, and poor quality) as the model. The reason why we added three quality levels for the expert summaries and crowdsourcing summaries was to explore whether different quality levels would yield different LSA cosine scores. Knowing their differences or similarities would help us interpret the results.

Efficient LSA method: crowdsourcing LSA

The results of Pearson correlations, comparisons of correlations with Fisher's z, and comparisons of the minimum and maximum correlations across the eight topics suggested that crowdsourcing LSA was an effective, efficient, and practical method to predict summaries as perceived by humans as well as expert summaries and could predict moderately better than the source text. This crowdsourcing LSA method does not need experts to pregrade or write students' summaries to generate a model summary and thus saves substantial human time, effort, and costs.

Olmos et al. (2011) reported that the expert summary method was better than the source text method and the individual summary method, with correlations with human scores ranging from .41 to .64 for the short summaries (25–50 words) of an expository text. Their results were similar to those in E. Kintsch et al.'s (2000) study. Our results were consistent with both studies and showed that the expert summary method was slightly better than the source text method, with correlations between expert summary LSA cosine scores and human scores ranging from .39 to .51 versus .29 to .48 between source text summary and human scores. Our results were slightly lower than those in previous studies, probably because we used the natural summaries that the participants wrote without any correction for spelling unlike in previous studies. The misspelt words may have led to a discrepancy between LSA scores and human scores. Misspelt words, however, will reduce quality of writing and are always included in the writing rubric (Friend, 2001). This is why we did not include spelling correction as the previous studies did. Another reason is that we only used one expert summary, whereas E. Kintsch et al. (2000) and Olmos et al. (2011) used four expert summaries. The more expert summaries that are included in the model summary, the more information that is provided in the model summary. Subsequently, the model summary may better capture the information provided in the target summary. In our study, the Crowdsourcing-Good method is equivalent to the multiple-

Table 6 Correlations of crowdsourcing LSA scores between different sample sizes of the model summary

Topic	Sample size	10	20	30	40	50	60	70	80	90	100
1	10	1.00									
	20	.97	1.00								
	30	.97	.99	1.00							
	40	.97	.98	1.00	1.00						
	50	.97	.98	.99	1.00	1.00					
	60	.97	.97	.99	1.00	1.00	1.00				
	70	.97	.97	.99	.99	1.00	1.00	1.00			
	80	.96	.96	.98	.99	.99	1.00	1.00	1.00		
	90	.95	.96	.98	.98	.99	.99	1.00	1.00	1.00	
	100	.95	.96	.98	.98	.99	.99	.99	1.00	1.00	1.00
2	10	1.00									
	20	.96	1.00								
	30	.89	.98	1.00							
	40	.88	.97	1.00	1.00						
	50	.89	.97	.99	1.00	1.00					
	60	.90	.97	.99	.99	1.00	1.00				
	70	.90	.97	.99	.99	.99	1.00	1.00			
	80	.89	.97	.98	.99	.99	1.00	1.00	1.00		
	90	.91	.97	.97	.97	.98	.99	1.00	1.00	1.00	
	100	.89	.96	.97	.97	.98	.99	1.00	1.00	1.00	1.00
3	10	1.00									
	20	.99	1.00								
	30	.97	.99	1.00							
	40	.97	.99	1.00	1.00						
	50	.97	.98	.99	1.00	1.00					
	60	.97	.98	.99	1.00	1.00	1.00				
	70	.96	.98	.99	.99	1.00	1.00	1.00			
	80	.96	.98	.99	.99	1.00	1.00	1.00	1.00		
	90	.95	.97	.99	.99	1.00	1.00	1.00	1.00	1.00	
	100	.95	.97	.98	.99	.99	1.00	1.00	1.00	1.00	1.00
4	10	1.00									
	20	.99	1.00								
	30	.97	.99	1.00							
	40	.94	.97	.99	1.00						
	50	.95	.97	.99	1.00	1.00					
	60	.96	.97	.99	1.00	1.00	1.00				
	70	.96	.97	.99	.99	1.00	1.00	1.00			
	80	.95	.97	.99	.99	1.00	1.00	1.00	1.00		
	90	.95	.96	.98	.99	1.00	1.00	1.00	1.00	1.00	
	100	.95	.96	.98	.99	.99	1.00	1.00	1.00	1.00	1.00
5	10	1.00									
	20	.97	1.00								
	30	.97	1.00	1.00							
	40	.98	.99	1.00	1.00						
	50	.97	.99	.99	1.00	1.00					
	60	.97	.99	.99	1.00	1.00	1.00				
	70	.96	.98	.99	1.00	1.00	1.00	1.00			
	80	.95	.98	.99	.99	1.00	1.00	1.00	1.00		
	90	.95	.98	.99	.99	.99	1.00	1.00	1.00	1.00	

Table 6 (continued)

Topic	Sample size	10	20	30	40	50	60	70	80	90	100
	100	.95	.98	.98	.99	.99	.99	1.00	1.00	1.00	1.00
6	10	1.00									
	20	.99	1.00								
	30	.99	1.00	1.00							
	40	.98	.99	1.00	1.00						
	50	.98	.99	1.00	1.00	1.00					
	60	.98	.99	1.00	1.00	1.00	1.00				
	70	.97	.99	1.00	1.00	1.00	1.00	1.00			
	80	.97	.99	.99	1.00	1.00	1.00	1.00	1.00		
	90	.97	.98	.99	1.00	1.00	1.00	1.00	1.00	1.00	
	100	.97	.98	.99	.99	1.00	1.00	1.00	1.00	1.00	1.00
7	10	1.00									
	20	.98	1.00								
	30	.97	1.00	1.00							
	40	.96	.99	1.00	1.00						
	50	.95	.99	1.00	1.00	1.00					
	60	.95	.99	.99	1.00	1.00	1.00				
	70	.94	.98	.99	1.00	1.00	1.00	1.00			
	80	.93	.98	.99	.99	1.00	1.00	1.00	1.00		
	90	.93	.97	.99	.99	.99	.99	1.00	1.00	1.00	
	100	.93	.97	.98	.99	.99	.99	1.00	1.00	1.00	1.00
8	10	1.00									
	20	.96	1.00								
	30	.94	.99	1.00							
	40	.92	.97	1.00	1.00						
	50	.93	.98	1.00	1.00	1.00					
	60	.92	.98	1.00	1.00	1.00	1.00				
	70	.92	.98	1.00	1.00	1.00	1.00	1.00			
	80	.91	.97	.99	1.00	1.00	1.00	1.00	1.00		
	90	.91	.97	.99	.99	.99	1.00	1.00	1.00	1.00	
	100	.91	.97	.99	.99	.99	1.00	1.00	1.00	1.00	1.00

All the correlations were significant, $p > .01$. Topic 1 = Butterfly and Moth; 2 = Hurricane; 3 = Walking and Running; 4 = Kobe and Jordan; 5 = Floods; 6 = Job Market; 7 = Effect of Exercising; 8 = Diabetes. 1–4 are the comparison texts; 5–8 are the causation texts

expert summary method. The disadvantage for more expert summaries, however, is that more human effort is needed.

Fortunately, crowdsourcing LSA predictions were comparable to the expert summary, with correlations ranging from .40 to .49 as well as to the Crowdsourcing-Good method (r = .41–.51). Theoretically, the Crowdsourcing-Good summary method should better predict human scores relative to the expert summary method because it included all the good summaries and maximally captured the content that good summaries should include. However, our study showed that the performance of the LSA method with many good summaries and with one good expert summary was not significantly different. It is likely that original source texts are short and experts can identify the important information without any disagreement. When the original source texts are long and experts have discrepancies in agreement regarding important information (E. Kintsch et al., 2000), the Crowdsourcing-Good method may exceed the expert-generated summary method.

On the other hand, Crowdsourcing LSA cosine scores had almost perfect correlations with LSA cosine scores of Crowdsourcing-Good summary. One possible reason is that the crowdsourcing summary contained all the summaries of the different quality levels. Similarly, the crowdsourcing good summary probably consisted of some unimportant or irrelevant content items or details that occurred in intermediate or poor crowdsourcing summaries. This was because good crowdsourcing summaries not only included perfectly written summaries scored as 8 points, but also included summaries

Table 7 Pearson correlations between the crowdsourcing summaries and human summary scores with increasing numbers of summaries in crowdsourcing summaries ($N = 80$)

Topic	10	20	30	40	50	60	70	80	90	100
1	.25*	.27*	.27*	.27*	.26*	.27*	.28*	.27*	.28*	.29**
2	.35**	.38**	.40**	.42**	.43**	.42**	.42**	.42**	.41**	.41**
3	.31**	.34**	.38**	.38**	.38**	.38**	.39**	.38**	.39**	.39**
4	.37**	.36**	.41**	.45**	.44**	.45**	.45**	.45**	.46**	.46**
5	.59**	.62**	.63**	.64**	.65**	.65**	.65**	.65**	.65**	.65**
6	.32**	.33**	.33**	.34**	.33**	.34**	.35**	.34**	.34**	.34**
7	.38**	.37**	.37**	.38**	.38**	.38**	.38**	.37**	.38**	.38**
8	.39**	.43*	.47**	.48**	.48**	.49**	.49**	.50**	.50**	.50**
Average	.37	.39	.41	.42	.42	.42	.42	.42	.43	.43

1 = Butterfly and Moth; 2 = Hurricane; 3 = Walking and Running; 4 = Kobe and Jordan; 5 = Floods; 6 = Job Market; 7 = Effect of Exercising; 8 = Diabetes. 1–4 are the comparison texts; 5–8 are the causation texts. * $p < .05$. ** $p < .01$

graded as 6 and 7 points. They included some unimportant or irrelevant information or details, a few of which occurred in intermediate summaries and most of which occurred in poor summaries. Mixing a huge amount of summaries may reduce the noisy interfering information, which would make LSA scores between the crowdsourcing method and the Crowdsourcing-Good method perfectly correlated. This further confirms that it is not necessary to pregrade summaries and select good summaries as the model summary. The simple, efficient, effective Crowdsourcing-All method explains an amount of variance in human summary scores comparable to those explained by the Crowdsourcing-Good summary and the expert summary, both of which involve great human effort.

The results also showed that the crowdsourcing method and Crowdsourcing-Good method moderately better predicted human scores relative to the Crowdsourcing-Intermediate method and Crowdsourcing-Poor method. One explanation is that the latter two summaries may have contained more content items that may not have occurred in good summaries, such as unimportant or irrelevant information or details. These findings imply that crowdsourcing summary should contain a certain number of good quality levels of summaries; otherwise, its predictability may slightly decrease. In the present study, the percentage of good summaries for each topic ranged from 25 to 50%. However, more empirical evidence is needed to confirm whether the increasing number of good summaries yields better performance.

Our results also showed that the crowdsourcing method, Crowdsourcing-Good method, and expert summary method more consistently predicted human scores than source text, but no significant differences were found. These results were different from the previous study (E. Kintsch et al., 2000; Olmos et al., 2011), in which source text was not a better method than expert summary. This difference may be caused by the different lengths of source texts. The source texts in our study were very short, about 300 words, three times longer than 100-word summaries. The short texts usually contained much more major information and less unimportant or irrelevant details, as compared to the longer source texts. Therefore, the shorter source texts may better predict content of summaries as perceived by humans. Even though Olmos et al. (2011) did not specify the length of the source text, more empirical evidence is needed to support this claim.

Olmos et al. (2011) actually used all the students' summaries as the model summaries, but they still compared the target summary with each individual summary and then assigned the average LSA cosine score to the target summary. They found that this method was not better than the pregraded summary or expert summary. It is likely that the average LSA cosine score represents the averaged information of all the individual student summaries. However, when we put all the summaries in one document, it will strengthen the useful information because the majority of good summaries contain similar useful information. The occurrence of the repeated useful information was given credit when the LSA vector was generated for the crowdsourcing summary because the vector had a larger component of information that was frequently repeated. This may explain why the crowdsourcing method could predict human scores as well as the expert summary method, whereas the average of the individual summary method could not.

Optimal number of crowdsourcing summaries

In the present study, we also compared the performance of the crowdsourcing methods with ten different numbers of crowdsourcing summaries, starting from 10 to 100 at intervals of 10, by adding ten additional summaries to each previous sample size. This method showed that the LSA scores from ten crowdsourced summaries were almost perfectly correlated with each other. It is possible that crowdsourcing LSA with ten summaries could reflect the majority of information that students constructed in their summaries. Adding more

summaries to the crowdsourcing summary beyond ten increased the content by only a negligible amount. Consequently, increasing the number did not improve the performance after this minimal number was reached.

These findings indicated that ten summaries could yield as good performance as of the large number of summaries, even though 40 or 90 summaries could slightly improve the performance. Therefore, this method could not only be used for large-scale, high-stake, and summative assessment, but also for small-scale, low-stake, and formative assessment. Specifically, one instructor who teaches a small class with about ten students is able to use the crowdsourcing LSA method to evaluate students' deep learning.

We noted some interesting results that appeared when we split the summaries into two corpora: summaries in the crowdsourcing corpus were only used as crowdsourcing summaries, whereas those in the target corpus were only used as the target summary to be scored. The correlations between crowdsourcing LSA and human scores had a larger variation across different topics. The highest correlation on the topic of *Flood* was .59 with ten crowdsourcing summaries and then increased to .65 with 100 crowdsourcing summaries, whereas the lowest correlation on the topic of *Butterfly and Moth* was as low as .25 and increased to .29 when more crowdsourcing summaries were included in the model summary. It is still unclear why splitting summaries into the crowdsourcing corpus and the target corpus leads to this discrepancy across topics. One possible explanation may be that the differences in sample sizes for the correlation analyses: 80 (after splitting corpus) versus 180. The large sample size may decrease the variation and lead to a more centralized result. However, further examination of summaries in the split corpora and overall summaries is needed to answer this question.

It is important to point out that AutoTutor CSAL enhanced participants' performance on summary writing and engagement, especially during training (four summaries) with scaffolding for the summarization strategy (Li & Graesser, 2017). The performance of crowdsourcing LSA, however, did not differ among the three quality levels (i.e., good, intermediate, and poor) of crowdsourcing summaries, which suggests that this method is independent of summary quality. Thus, the crowdsourcing LSA method can be generalized to summaries written with other computer-assisted systems or only in a word processor.

Conclusions and future directions

In the present study we proposed a crowdsourcing-based LSA method to predict human summary scores and then evaluated this method by comparing it with different methods, including expert summary and the source text. The results showed that the crowdsourcing LSA method is a promising approach, as it predicted human scores as well as the expert summary did, and slightly better than the source text. Further analyses on the number of crowdsourcing summaries showed that the number had no apparent effect on performance: The crowdsourcing LSA method using ten summaries could predict human scores as well as the method with 100 summaries.

The study has advanced the research on automated summary scoring in two ways. First, this study proposes a method for computerized summary scoring without the involvement of human experts to generate the model summary, and the performance of this method was comparable to that based on an expert summary, which has been considered the most efficient and practical method to date. Crowdsourcing LSA saves abundant human effort and time, and it is a promising method for CSS in the long run. Moreover, this study has shown that good performance can be achieved with a minimal number of ten crowdsourcing summaries, which is an acceptable number for school teachers in their routine instruction. To sum up, the crowdsourcing LSA approach may advance research on automated assessment and also motivate teachers to use summary writing to foster deep learning, as this method demands no human effort to write or pregrade summaries but yields as good performance as the expert-summary method, which does require human effort.

However, we should mention some limitations of this study. First, the original source texts were short, about 300 words, whereas in previous studies, the source texts usually had a thousand of words apiece. In fact, summarizing a longer passage or several passages on the same topic is a common task for higher grade students. Therefore, it will be necessary to replicate this study with longer source texts. Second, a future study might include more texts so that we could investigate whether the variance among texts in their correlations between the crowdsourcing method and human scores is itself correlated with features of the particular texts, such as formality, FKGL, and word count. Third, a future study might examine whether text structure is related to the performance of the crowdsourcing approach. For example, is the crowdsourcing method best suitable to a compare-and-contrast, cause-and-effect, problem-and-solution, sequence, or description structure? Finally, it is uncertain whether the proportions of good, intermediate, and poor summaries affect performance. The present study only showed that varied proportions did not affect performance across different topics. However, more empirical evidence for proportions of different levels of quality of summaries within each topic will be needed to make solid confirmation of this claim.

Because this study has concentrated on the evaluation of the crowdsourcing LSA method, there is no space to report the summary grades that the crowdsourcing LSA method computed. A future study will focus on summary grading, using the crowdsourcing LSA method alone or along with other

linguistic and discourse features. Olmos et al. (2011) reported that a new LSA algorithm—a best-dimension algorithm—improved LSA performance with an expert summary as the model summary. It will be worth examining whether this will also be true using the crowdsourcing summary as the model summary, by comparing a best-dimension algorithm with the standard LSA algorithm.

A summary is a condensed form of an original source text and a good summary typically reflects the important information in the source text. Therefore, good summaries tend to point to similar important pieces of information stated in the original source texts. The methodology can be expanded to other automated assessments in which the student responses are expected to semantically converge on subject matter content, such as scientific explanations in science education in the format of open-ended questions. For example, scientific explanations for science inquiry require that a student states a claim and provides evidence to support the claim with reasoning. Because the content for scientific explanations is fixed, it is worth investigating whether this method works for scientific explanation or other convergent questions. Differently, essays do not necessarily contain specific, convergent content. Writers choose diverse information to support their arguments, which makes content diverse. Therefore, this method may not be appropriate for essays. However, evidence is needed to support this claim.

To sum up, the crowdsourcing LSA method is a promising approach to the automated assessment of summaries. After a tool for automated summary assessment has been developed, the application of this method will promote school teachers to use the summarization strategy in their daily instruction, and ultimately enhance the instruction of deep comprehension and learning.

Author note This work was funded by the Institute of Education Sciences (Grant No. R305C120001). Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of these funding agencies, cooperating institutions, or other individuals.

References

Baleghizadeh, S., & Babapur, M. (2011). The effect of summary writing on reading comprehension and recall of EFL students. *New English Reading Association Journal*, *47*, 44–57.

Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology*, *25*, 313–339.

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

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. https://doi.org/10.1177/1745691610393980

Carroll, D. W. (2008). Brief report: A simple stimulus for student writing and learning in the introductory psychology course. *North American Journal of Psychology*, *10*, 159–164.

Chiu, C. H. (2015). Enhancing reading comprehension and summarization abilities of EFL learners through online summarization practice. *Journal of Language Teaching and Learning*, *5*, 79–95.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*, 197–202. https://doi.org/10.3758/BF03204765

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111–128.

Friend, R. (2001). Effects of strategy instruction on summary writing of college students. *Contemporary Educational Psychology*, *26*, 3–24.

Garner, R., & McCaleb, J. L. (1985). Effects of text manipulations on quality of written summaries. *Contemporary Educational Psychology*, *10*, 139–149.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology*, *35*, 157–173.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elementary School Journal*, *115*, 210–229.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193–202. https://doi.org/10.3758/BF03195564

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology*, *28*, 597–606.

Hu, X., Cai, Z., Wiemer-Hastings, Graesser, A. C., & McNamara, D. S. (2007). Strengths, limitations, and extensions of LSA. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch. (Eds.), *The handbook of latent semantic analysis* (pp. 401–426). Mahwah, NJ: Erlbaum.

Jitendra, A., Cole, C., Hoppes, M., & Wilson, B. (1998). Effects of a direct instruction main idea summarization program and self-monitoring on reading comprehension of middle school students with learning disabilities. *Reading and Writing Quarterly*, *14*, 379–396.

Jitendra, A., Hoppes, M., & Xin, Y. (2000). Enhancing main idea comprehension for students with learning problems: The role of a summarization strategy and self-monitoring instruction. *Journal of Special Education*, *34*, 127–139.

Jorge-Botana, G., Luzón, J. M., Gómez-Veiga, I., & Martín-Cordero, J. I. (2015). Automated LSA assessment of summaries in distance education some variables to be considered. *Journal of Educational Computing Research*, *52*, 341–364.

Karpicke, J. D., & Roediger, H. L., III (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction*, *7*, 161–195.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R., & LSA Research Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Environments*, *8*, 87–109.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202. https://doi.org/10.1207/s15516709cog2502_1

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157–170). Amsterdam, The Netherlands: Benjamins.

Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly*, *10*, 62–102.

Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic Analysis theory. *Current Directions in Psychological Science*, *7*, 161–164.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice*, *48*, 44–52.

Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods*, *38*, 616–627. https://doi.org/10.3758/BF03193894

Leopold, C., Sumfleth, E., & Leutner, D. (2013). Learning with summaries: Effects of representation mode and type of learning activity on comprehension and transfer. *Learning and Instruction*, *27*, 40–49.

Li, H., Cai, Z., & Graesser, A. C. (2016). How good is popularity? Summary grading in crowdsourcing. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 430–435). Raleigh, NC: EDM Society.

Li, H., Cheng, Q., Yu, Q., & Graesser, A. C. (2015). The role of peer agent's learning competency in trialogue-based reading intelligent systems. In C. Conati & N. T. Hefferman (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (pp. 694–697). Berlin, Germany: Springer.

Li, H., & Graesser, A. C. (2017). Impact of pedagogical agents' conversational formality on learning and engagement. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education: AIED 2017* (Lecture Notes in Computer Science, Vol. 10331, pp. 188–200). Beijing, China: Springer.

Li, H., Graesser, A. C., & Cai, Z. (2013). Comparing two measures of formality. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference* (pp. 220–225). Palo Alto: AAAI Press.

Li, H., Shubeck, K., & Graesser, A. C. (2016). Using technology in language assessment. In D. Tsagari & J. V. Banerjee (Eds.), *Contemporary second language assessment: Contemporary applied linguistics* (Vol. 4, pp. 281–297). London, UK: Bloomsbury Academic.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In B. Webber & D. Byron (Eds.), *Text summarization branches out: Proceedings of the 2004 ACL Workshop on Discourse Annotation* (Vol. 8, pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Lin, C. Y., & Hovy, E. (2003). Automatic evaluation of summaries using *n*-gram co-occurrence statistics. In M. Heartst & M. Ostendorf (Eds.), *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (Vol. 1, pp. 71–78). Edmonton, Canada: Association for Computational Linguistics.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*, 57–86.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science*, *44*, 567–581.

Nenkova, A., & Passonneau, R. J. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In S. Dumais, D. Marcu, & S. Roukos (Eds.), *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting* (Vol. 4, pp. 145–152). Boston, MA: Association for Computational Linguistics.

Oded, B., & Walters, J. (2001). Deeper processing for better EFL reading comprehension. *System*, *29*, 357–370.

Olmos, R., León, J. A., Escudero, I., & Jorge-Botana, G. (2011). Using latent semantic analysis to grade brief summaries: Some proposals. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*, 192–209.

Olmos, R., León, J. A., Jorge-Botana, G., & Escudero, I. (2009). New algorithms assessing short summaries in expository texts using latent semantic analysis. *Behavior Research Methods*, *41*, 944–950. https://doi.org/10.3758/BRM.41.3.944

Olmos, R., León, J. A., Jorge-Botana, G., & Escudero, I. (2013). Using latent semantic analysis to grade brief summaries: A study exploring texts at different academic levels. *Literary and Linguistic Computing*, *28*, 388–403.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188. https://doi.org/10.1177/0963721414531598

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In P. Isabelle (Ed.), *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Philadelphia, PA: Association for Computational Linguistics.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015* (UT Faculty/Researcher Working Paper). Department of Psychology, University of Texas, Austin, TX.

Perin, D., & Lauterbach, M. (2016). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*. Advance online publication. https://doi.org/10.1007/s40593-016-0122-z

Perin, D., Lauterbach, M., Raufman, J., & Kalamkarian, H. S. (2016). Text-based writing of low-skilled postsecondary students: Relation to comprehension, self-efficacy and teacher judgments. *Reading and Writing*. Advance online publication. https://doi.org/10.1007/s11145-016-9706-0

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337–354.

Rogevich, M., & Perin, D. (2008). Effects on science summarization of a reading comprehension intervention for adolescents with behavior and attention disorders. *Exceptional Children*, *74*, 135–154.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*, 213–220. https://doi.org/10.1177/2167702612469015

Shokrpour, N., Sadeghi, A., & Seddigh, F. (2013). The effect of summary writing as a critical reading strategy on reading comprehension of Iranian EFL learners. *Journal of Studies in Education*, *3*, 127–138. https://doi.org/10.5296/jse.v3i2.2644

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Spirgel, A. S., & Delaney, P. F. (2016). Does writing summaries improve memory for text? *Educational Psychology Review*, *28*, 171–196.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*, 155–167. https://doi.org/10.3758/s13428-010-0039-7

Stewart, T. L., Myers, A. C., & Culley, M. R. (2010). Enhanced learning and retention through "writing to learn" in the psychology classroom. *Teaching of Psychology*, *37*, 46–49.

Sung, Y.-T., Liao, C.-N., Chang, T.-H., Chen, C.-L., & Chang, K.-E. (2016). The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique. *Computers & Education*, *95*, 1–18. https://doi.org/10.1016/j.compedu.2015.12.003

Trabasso, T., & Bouchard, E. (2002). Teaching readers how to comprehend texts strategically. In C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 176–200). New York, NY: Guilford Press.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension* (pp. 11–12). New York, NY: Academic Press.

Wade-Stein, D., & Kintsch, E. (2004) Summary Street: Interactive computer support for writing. *Cognition and Instruction*, *22*, 333–362.

Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders*, *30*, 275–287.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. P. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 535–542). Amsterdam, The Netherland: IOS Press.