Int J Artif Intell Educ (2018) 28:29–55
DOI 10.1007/s40593-016-0128-6

ARTICLE

Wise Crowd Content Assessment and Educational Rubrics

Rebecca J. Passonneau[1] · Ananya Poddar[1] · Gaurav Gite[1] · Alisa Krivokapic[1] · Qian Yang[2] · Dolores Perin[3]

Published online: 5 December 2016

Abstract Development of reliable rubrics for educational intervention studies that address reading and writing skills is labor-intensive, and could benefit from an automated approach. We compare a main ideas rubric used in a successful writing intervention study to a highly reliable wise-crowd content assessment method developed to evaluate machine-generated summaries. The ideas in the educational rubric were extracted from a source text that students were asked to summarize. The wise-crowd content assessment model is derived from summaries written by an independent group of proficient students who read the same source text, and followed the same instructions to write their summaries. The resulting content model includes a ranking over the derived content units. All main ideas in the rubric appear

✉ Rebecca J. Passonneau
becky@ccls.columbia.edu

Ananya Poddar
ap3317@columbia.edu

Gaurav Gite
gg2614@columbia.edu

Alisa Krivokapic
ak3533@columbia.edu

Qian Yang
laraqianyang@gmail.com

Dolores Perin
perin@tc.edu

[1] Columbia University, New York, NY, USA

[2] Tsinghua University, Beijing, China

[3] Teachers College of Columbia University, New York, NY, USA

prominently in the wise-crowd content model. We present two methods that automate the content assessment. Scores based on the wise-crowd content assessment, both manual and automated, have high correlations with the main ideas rubric. The automated content assessment methods have several advantages over related methods, including high correlations with corresponding manual scores, a need for only half a dozen models instead of hundreds, and interpretable scores that independently assess content quality and coverage.

Keywords Automated content analysis · Writing intervention · Wise-crowd content assessment · Writing rubrics

Introduction

Automated tools to identify strengths and weaknesses of students' reading and writing skills could make it easier for teachers across disciplines to promote reading skills, and to incorporate more writing into their curricula. Of the many aspects of verbal skills that students need to learn, this paper focuses on the assessment of their mastery of content. We present a method to assess content of students' written summaries that derives a model of the important content for a particular summarization task from a small set of examples. For reasons explained below, we refer to the authors of the example set as a *wise crowd*. We demonstrate the application of the wise-crowd method on summaries written by community college students who participated in a successful intervention study to improve reading and writing skills. We present a main ideas rubric used in the community college intervention study, followed by a description of wise-crowd content assessment, and two automated implementations of the method. Experimental results from a comparison of scores based on the rubric for a sample of 120 student summaries correlate highly with manual and automated applications of the wise-crowd content assessment.

The work presented here is significant because of the importance of reading and writing skills, alongside serious deficiencies in students' acquisition of these skills. It is widely acknowledged that in today's information society, reading is a critical skill. Writing skills are important in many professions (Burstein et al. 2016); for the generation of scientific knowledge (Norris and Phillips 2003; Yore et al. 2004); to demonstrate students' mastery of content (Graham and Perin 2007b); to foster learning (Bangert-Drowns et al. 2004; Sampson et al. 2013; Hand et al. 2004). Reading, and especially writing, occupy a central place in national standards, as in the 2010 Common Core State Standards (CCSS) for elementary and secondary education. For at least a decade, however, a majority of students in our educational system have not been achieving grade level proficiency in reading or writing, as documented in reports from the National Center for Education Statistics (Persky et al. 2003; Salahu-Din et al. 2008; NCES 2012; Glymph 2010; 2013; Glymph and Burg 2013). Studies of the development of writing skills point to a very long time course (Kellogg 2008), interdependence with other cognitive skills (Deane et al. 2008), the benefits of large amounts of practice and feedback, and the need for explicit guidance in the demands

of different genres (Beers and Nagy 2011; 2009; Olinghouse et al. 2015; Olinghouse and Wilson 2013). Results from recent surveys of high school and middle school teachers, however, indicate that teachers of subjects other than language arts feel ill-prepared to teach writing, and devote insufficient classroom time to writing (Graham et al. 2014; Gillespie et al. 2014).

Assessment of student writing for content plays a key role in educational interventions for reading and writing skills, and to assess mastery of disciplinary knowledge. Due to the expressive variation language affords, and the judgment required to determine if a student has articulated a given idea, specification of a content rubric and evaluation of its interrater reliability are essential in such studies. The rubrics, however, are time-consuming to develop and administer, and differences in rubrics across studies impede direct comparison of results. Automated rubrics could facilitate such studies, and potentially lead to standardization. Optimism about this possibility is justified by strong analogs with annotation efforts in the field of Natural Language Processing (NLP), where one goal is to develop automated methods to analyze text. NLP annotation schemes vary widely, and can capture anything from parts of speech (noun, verb, adjective, etc.) to the argument structure of a text. An NLP annotation scheme is analogous to a coding rubric in that it consists of instructions for people to label portions of a text or transcript. An annotation scheme is then evaluated by measuring the inter-annotator reliability, analogous to interrater reliability of an assessment rubric in education research. Annotated data is then used to develop, train and test NLP methods to automatically replicate the manual annotation. The parallels between education rubrics and NLP annotation tasks argue for the potential benefit of investigating whether the rubrics could be automated. The present paper makes a step in that direction through its comparison of an NLP content assessment method with an intervention rubric for reading and writing skills.

Two works that apply NLP methods to automate rubrics designed to trigger feedback on the content of students' writing rely on domain experts to create lists of concepts, and on large numbers of models to train automated methods (Gerard et al. 2016; Rosé and VanLehn 2005). Instead of reliance on domain experts to specify the content, we use a *wisdom-of-the-crowd* approach (Surowiecki 2004), meaning we aggregate a content model from the summaries of multiple individuals. Our earlier work applied this approach through manual methods to assess content selection in news summaries generated by machines (Nenkova and Passonneau 2004; Nenkova et al. 2007). Manual application of this method to a small sample of student essays was found to correlate highly with the intervention rubric mentioned above (Pearson = 0.89; corrected from 0.85) (Passonneau et al. 2013). Experiments to automate the assessment step, given an existing, manually generated wise-crowd content model, correlated very highly with the manual version (Pearson = 0.93 $\pm$ 0.01). This paper extends that study to a much larger sample of student essays, describes an enhancement to the automated assessment, and an alternative method that automates both the content model generation and the assessment (Yang et al. 2016).

The manual wise-crowd content assessment, and both automated methods, have high correlations with the educational rubric. The strong results suggest that the wise-crowd content method could be substituted for rubrics that assess content, or

supplement rubrics that incorporate content along with other dimensions. Advantages to automation of rubrics include reduction of effort, and comparability across different studies. Advantages of the wise-crowd method over other automated approaches to assessment are that it requires a small sample of training examples rather than hundreds or thousands, and it provides interpretable scores that give fine-grained measures of content quality and coverage.

Related Work

This section covers evaluation of automated summarizers in NLP, and relevant work from Automated Essay Scoring (AES). Automatic summarizers can take in large volumes of electronic text and produce condensed summaries of a desired length. Two wise-crowd methods to evaluate the content of machine-generated summaries relative to human ones were developed concurrently (Nenkova and Passonneau 2004; Teufel and van Halteren 2004). The method used here is distinctive in explicit assignment of weights to Content Units (CUs). It was employed in two large-scale evaluations of automatic summarizers conducted by the National Institute of Standards (NIST) (Passonneau et al. 2005; Passonneau et al. 2006). Automated methods for content evaluation of machine-generated summaries are more prevalent, but are not a good fit for student essays, due to inaccuracy on individual summaries. AES methods have become prevalent for student essay scoring (Page 1994; Foltz et al. 1999; Burstein et al. 1998; Attali and Burstein 2006; Rudner et al. 2006; Shermis and Burstein 2013; Kharkwal and Muresan 2014), and have also been incorporated in computer-based learning (CBL) environments (Proske et al. 2012; Roscoe and McNamara 2013; Roscoe et al. 2015b; Roscoe et al. 2015a; Gerard et al. 2016). In contrast to AES, the wise-crowd method specifically assesses content, and has higher reliability.

The challenge in evaluating summary content is that no two people will include the same information in their summaries. The wise-crowd content evaluation methods mentioned above were proposed concurrent with publication of *The Wisdom of Crowds* (Surowiecki 2004). It had long been observed that for certain questions, especially quantitative ones, a more confident solution can be assembled from many answers, as opposed to reliance on a single response. Surowiecki points to a 1907 *Nature* article by Francis Galton as the beginning of a precise formulation of the wise-crowd concept. Galton showed that the median of hundreds of individuals' guesses of the weight of a particular ox was within 1 % of its true weight, while any one guess would have a high probability of being very far off. As discussed below, we have shown that wise-crowd content assessment is reliable with about five wise-crowd members.

Surowiecki summarized the properties of a wise crowd as independence of individuals' judgments, diversity of individuals, equal access to knowledge, and a method to aggregate the answers. For subjective questions, aggregation of the crowds' responses is problematic. Pyramid annotation (Nenkova and Passonneau 2004) and factoid annotation (Teufel and van Halteren 2004) independently provide a wise-crowd approach to content evaluation of summaries, based on similar methods to

aggregate linguistic meaning, which is notoriously subjective. Both kinds of annotation apply to human summaries selected to serve as models. Both define meaning units (content units, or factoids) to represent semantic overlap among model summaries, and both require the human annotator to construct a phrase to represent the shared semantics. This descriptive phrase constitutes the factoid in Teufel and van Halteren (2004). In pyramid annotation, however, annotators must also select no more than one *contributor* phrase from each summary that expresses the same content. The size of the set of contributors serves as the weight of a content unit (CU), as explained in a later section (see Fig. 1). Weights differentiate CUs by their importance, and support quantitative evaluation. Other advantages to the pyramid method are that the annotation procedure is simple and easy to learn, as evidenced by its use for large-scale evaluations. As discussed below, it has been subjected to rigorous evaluation of annotation reliability and score stability.

The weights of CUs are an emergent phenomenon, meaning they can be observed only given a wise crowd. The rank of CUs by weight has been observed to follow a power law distribution (Nenkova and Passonneau 2004; Teufel and van Halteren 2004; Qazvinian and Radev 2012), as in word frequency (Zipf 1949) or sense frequency (Passonneau et al. 2012). A few content units occur in almost all models (high weight), an intermediate number occur in some models (moderate weight), and a very long tail of content units occur in only a few (low weight). In the following section we discuss how this differentiation of content by weight aligns well with assessment rubrics for summarization of source texts to assess how well students can select important content (Day 1986; Perin et al. 2013).

Automatic evaluation of content selection facilitates progress and standardizes results in machine summarization. The most prevalent tool to automatically score machine-generated summaries is ROUGE (Lin 2004). Like the pyramid method, it uses model summaries as references to match all substrings in a target summary with all the substrings in the models. The score is a count of matched substrings normalized by total substrings (Lin 2004). Depending on the settings and summarization task, ROUGE scores often correlate highly with human responsiveness scores (Pearson ≥ 0.90) on macro-level evaluation. Responsiveness is a holistic score on a 5-point scale assigned by human asessors that combines content selection and quality of presentation (Owczarzak et al. 2012). Macro-level evaluation ranks systems by their average performance on multiple summarization tasks. The micro-level assesses a single summarization task. Louis and Nenkova (2009) observe that ROUGE performs poorly on micro-level evaluation. They provide a scoring method that avoids reliance on human models through comparison of word frequency distributions between summaries and source texts. Their model-free approach also does well for macro-level scoring, but poorly on micro-level evaluation (Saggion et al. 2010; Louis and Nenkova 2013). These methods would not do well in an educational setting, where micro-level scoring of individual essays is the norm. The results presented here show the reliability of the wise-crowd method for the macro-level; micro-level reliability is presented in the section entitled *Wise Crowd Content Assessment*.

The most prevalent NLP technology in educational assessment is AES, and most AES relies on one of two contrasting approaches. The analytic approach makes use

of observable linguistic properties of text, such as vocabulary, sentence complexity, and discourse structure (Burstein et al. 1998; Attali and Burstein 2006; Rudner et al. 2006; Shermis and Burstein 2013; Kharkwal and Muresan 2014). Statistical approaches to meaning, such as Latent Semantic Analysis (LSA), rely on word distributions to infer hidden meaning (Foltz et al. 1999; Wiemer-Hastings et al. 1999; Hughes et al. 2012). Synonyms can occur in the same sentences, thus meaning can be represented by the frequency of words in different contexts. Frequency matrices of words by contexts can be quite large, so methods like singular value decomposition are used to reduce the dimensionality, which also reduces noise. AES systems are typically trained to replicate manually assigned holistic quality scores of student essays on 3- or 4-point scales. Training for either method can require thousands of graded essays (Page 1994; Foltz et al. 1999; Burstein et al. 1998; Attali and Burstein 2006; Rudner et al. 2006). AES scores have been shown to correlate with manual scores as well as human raters correlate with each other. For example, the percent agreement between the e-rater system and human scores, and of human essay scorers with each other, are both in the range of about 45 % to 60 %, depending on the question prompt (Attali and Burstein 2006).

AES has been incorporated in various kinds of instructional environments. Intelligent Tutoring Systems (ITS) mainly focus on objective learning tasks (Roscoe and McNamara 2013), but there have been a few ITS systems for writing, such as Escribo (Proske et al. 2012) and Writing Pal (Roscoe and McNamara 2013; Roscoe et al. 2015a; Roscoe et al. 2015b). These ITS systems depend on AES methods to adapt system actions to the performance of a given student. Recent work argues for more focused approaches to automated assessment that would address specific aspects of writing (Deane 2013; Burstein et al. 2016). Beigman-Klebanov et al. (2014), for example, explore the use of content importance models (inspired in part by the pyramid method) to score students' use of sources. Application to other kinds of writing tasks includes integration of ideas in history (Hughes et al. 2012), relating STEM material to daily life (Beigman-Klebanov 2015), and scoring concepts in students' science writing (Liu et al. 2014).

Apart from occasional early work such as Butcher and Kintsch (2001), there have been few attempts to automate rubrics for writing interventions. Two recent attempts were applied to students' science explanations (Rosé and VanLehn 2005; Gerard et al. 2016). Explanation is a central aspect of science learning (Berland and McNeill 2010; Reiser and Kenyon 2012; Klein and Rose 2010), and evaluating explanations necessarily depends in part on assessment of conceptual content. Why2 was a tutoring system for students' written explanations of conceptual physics (VanLehn et al. 2002). For a given question prompt, such as where and why an object will fall when it is thrown by someone running, experts first established a fixed set of reasoning steps. Analytic and statistical methods were compared for the task of automatic classification of sentences from students' essays into one of the reasoning steps. A hybrid approach that combined statistical and analytic methods performed best.

WISE is a web-based platform to design and deliver science inquiry activities that has served over 100,000 K-12 students (Slotta and Linn 2009). In Gerard et al. (2016), the c-rater system (Leacock and Chodorow 2003) was adapted for WISE curricula to automatically identify pre-determined concepts in middle school students' short,

written answers. The automated scores triggered feedback for student revisions. The c-rater agreement with expert human scoring was 59 % compared with 62 % between two trained human raters (Liu et al. 2014). Compared with expert human assignment of Knowledge Integration (KI) guidance, c-rater and teacher-selected KI guidance were equally accurate. Students showed significant improvements with human or automated guidance, in contrast to generic guidance that might advise a student to reread source texts. Percent agreement, used by Liu et al. (2014), is an upper bound for the chance-adjusted agreement measures we use, because the latter factor out the proportion of agreements expected by chance (Passonneau and Carpenter 2014). Thus the high chance-adjusted agreement we find for wise-crowd content annotation corresponds to comparatively higher reliability. As reported in the section on Wise-Crowd Content Asessment, content models for ten summarization tasks had chance-adjusted agreement ranging from 0.68 to 0.89, and annotation of new summaries on three tasks had average chance-adjusted agreement of 0.78. Further, the automated wise-crowd method needs only four to five models instead of hundreds.

Summarization Rubrics for Educational Interventions

Summarization has long been viewed as an important skill to demonstrate reading comprehension of text (van Dijk and Kintsch 1977; Brown and Day 1983). Explicit instruction in summarization has been found to have a strong impact on students' reading and writing skills (Graham and Perin 2007a). This section points out two parallels of the wise-crowd method with rubrics from early research to understand students' ability to summarize what they have read: identification of proposition-like units in the source texts, and their differentiation by importance (Brown and Day 1983; Brown et al. 1983; Johnson 1970; Garner 1985; Day 1986; Turner 1987; Westby et al. 2010). Idea units are an analog to CUs, and importance ranking is an analog to CU weights, with the difference that CUs are derived from model summaries rather than source texts. The rubrics are generally arrived at and validated through consensus among large numbers of raters, rather than testing interrater reliability.

Kintsch and van Dijk (1978) developed an influential theory of the cognitive skills required to understand and produce text, as reflected in summaries. In their model, readers' comprehension relies on macroprocesses that transform the fine-grained propositional content of a text into a mental model (situational model) that depends on the reader's ability to generalize and draw inferences (van Dijk and Kintsch 1977; Kintsch and van Dijk 1978; van Dijk and Kintsch 1983). Brown and Day (1983) posit macrorules to codify the theory of van Dijk and Kintsch: omit irrelevant information, generalize and integrate specific information into more concise propositions, and synthesize novel propositions based on inferences. To investigate the acquisition of summarization skills in students from 5th grade through college, they developed a rubric to compare the content of summaries from students at different age levels, rather than to assess content per se. The rubric consisted of important idea units where one set of raters segmented the source texts into idea units, and a second set of

raters judged their importance in the text on a 4-point scale. Idea units that had average ratings of 3 or greater were taken as important. A related study of the influence of students' planning activities (drafts, note-taking) (Brown et al. 1983) coded idea importance using a method proposed in (Johnson 1970). One set of raters divided source texts into idea units using a criterion of where pauses can occur, then three new sets of raters eliminated either one-quarter, one-half or three-quarters of the sentences. Sentences that were never eliminated were taken to express important content. This method was also used in a study to investigate reasons for poor summarization (Garner 1985). In a study comparing alternative instruction methods for summarization, raters judged the importance of each sentence in a source text (Day 1986). A study of the inferential processes that support generalization and integration relied on a semi-automated method to identify propositions in a text (Turner 1987), then differentiated them by their role in the text (Turner and Greene 1978).

More recent work on summarization skills incorporates content as one dimension of a rubric, or adds relations among idea units. A study of instructional methods for summarization used a rubric of six dimensions, one of which addressed the quantity, accuracy and relevance of content (Westby et al. 2010). Another recent study of students' integration of material from multiple source texts compared students who wrote summaries to those who wrote argumentative essays (Gil et al. 2010). A single rubric for both types of writing was developed that first segmented the students' summaries and essays, rather than the source texts, into idea units, following Magliano et al. (1999). Idea units were then coded for one of four relationships to the source text, based on Wiley and Voss (1996): paraphrase, elaboration, addition, and misconception, with percent agreement of 85 %.

The studies cited above differ in their aims. They share, however, the goal of specifying a set of ideas to represent comprehension and integration of source material. In contrast to Brown and Day (1983) and similar work, a wise-crowd content model is generated from model summaries, rather than from the source texts. Here, the students to be assessed attend a community college to get additional preparation for postsecondary education. Therefore, as described below, the wise-crowd members were students who already perform at a level the community college students aim to achieve. In contrast to Gil et al. (2010), we compare the ideas expressed in a student's summary to the aggregation of ideas expressed in model summaries, rather than directly to the source. Further, the wise-crowd method makes no differentiation among ideas that paraphrase, elaborate, add to or misrepresent the source. It is possible, however, to add relations among content units found through the wise-crowd method, as in a study of children's oral narrative skills that identified narrative relations among content units (Passonneau et al. 2007).

A Rubric for Contextualized Curricular Support

A recent study conducted by one of the co-authors investigated reading and writing interventions for community college students, and found positive effects on several measures of written summarization for science versus generic texts (Perin et al. 2013). The rubric for the summarization tasks, like the rubrics for summarization

skills discussed above, identified important ideas in the source text. In a pilot study, we compared that rubric with wise-crowd content assessment for a sample of twenty students on a summarization task we refer to here as the *What is Matter?* task (Passonneau et al. 2013). The similarity in the goals of the two content assessment methods, and a very high Pearson correlation between the rubric and the manual wise-crowd content assessment, encouraged us to extend the pilot study. In the study presented here, the original twenty summaries are used to tune parameters for the automated wise-crowd methods, which are then tested on a new, larger set of summaries from 120 new students. Here, we briefly describe the intervention study and the rubric for the *What is Matter?* task.

The intervention study in Perin et al. (2013) tested the hypothesis that contextualized instruction could help academically unprepared students improve their reading and writing skills. Contextualized instruction, also referred to as embedded, anchored, integrative, theme-based or infused, focuses "teaching and learning squarely on concrete applications in a specific context that is of interest to the student" (Mazzeo et al. 2003). Among other questions, the study investigated whether reading comprehension and written summarization skills would improve after summarization practice, defining vocabulary terms, answering reading comprehension questions, and writing opinion essays.

A sample of 16 developmental education classrooms was recruited, with 12 classrooms receiving the intervention and the others serving as a comparison group. Mean reading scores were at the 22nd percentile for the Nelson-Denny Reading Test and at the 27th percentile for the Woodcock-Johnson Writing Fluency subtest. Different source texts were used over a 10-week period. Intact passages from both middle school and introductory college textbooks were used. The middle school text *What is Matter?* was used to build background knowledge and support the comprehension of subsequently-presented college level text. After reading the text passage, students answered a series of questions that focused on the main ideas of the passage. They then wrote a summary, in response to the following instruction:

> Write a summary of one or two paragraphs. A summary is a statement mostly in your own words that contains the important information in the passage. Please write clearly!

Summaries were scored on several dimensions, including the representation of main ideas from the source text. Fourteen main ideas in the source text were identified by a panel of three researchers who first worked independently, then came to a consensus on the final list. Main ideas in students' summaries were counted, and reported as a proportion of the total. Interrater reliability on a random selection of 25 % of the writing summaries had a Pearson correlation of 0.92.[1] In the next section, we describe the creation of a wise-crowd model for the same text, and compare it with the main ideas rubric.

[1] Personal communication with Perin.

Wise-Crowd Content Assessment

This section explains the annotation method to create a wise-crowd content model from model summaries, and the use of the content model to score new summaries. It also summarizes our previous work on the reliability and validity of the approach, originally referred to as the pyramid method (Nenkova and Passonneau 2004; Nenkova et al. 2007; Passonneau 2010).

In an educational setting, student learning can be scaffolded through examples that are somewhat beyond their skill level, the Vygotskyan zone of proximal development. Adopting this perspective, we collected model summaries from five masters students whose grades were high, and who were in the second semester of a highly competitive private northeastern university. We assume that the high admission standards of the school and the scholastic achievements of the students are indicators that their reading and writing proficiency would be more advanced than that of the students at the community college who participated in the intervention study. The comparison is analogous to the findings in Brown and Day (1983) that masters students in English had stronger summarization skills than community college students, but still had room for improvement. The wise-crowd students read the same source text as the community college students, answered the same reading questions, and followed the same instructions. The student co-author of (Passonneau et al. 2013) annotated the CUs in these model summaries after being trained by the first author, using guidelines created for large scale evaluations conducted by NIST (see Related Work section).[2] An annotation tool available from the guidelines website was used.

As briefly described in the Related Work section, Content Units (CUs) annotated in model summaries have three components: a set of contributing phrases from the model summaries that express the same meaning, a descriptive label composed by the annotator, and the cardinality of the contributors. The latter serves as a weight reflecting the importance of the CU. Figure 1 shows two CUs of weight 4 and two of weight 1 identified in the model summaries for the *What is Matter?* task. CU annotation is an iterative process in which the annotator looks for phrases across model summaries that express the same meaning, and selects them as contributors to the same CU. A model summary can contribute to each CU at most once. Summaries generally avoid redundancy, but if an idea is expressed again in the same model summary, the annotator is free to select either or both as the single contributor. The annotator provides a descriptive label for a CU that expresses the shared meaning, as illustrated in boldface for the four CUs in Fig. 1. CU105 in Fig. 1 has four contributors, one each from the first, second, third and fifth model summaries, which gives it a weight of four (boldface W=<N>). For clarity, the figure shows the contributor phrases in brackets, along with the remainder of the original sentential context in italics; the italicized text

[2]The guidelines at http://www1.ccls.columbia.edu/beck/DUC2006/2006-pyramid-guidelines.html were prepared for the 2006 Document Understanding Conference organized by NIST. Designers of approximately two dozen systems participated in the 2006 evaluation (Passonneau et al. 2006).

MODEL: Four Content Units (CUs)

CU105 W=4	**Matter is what makes up all objects or substances.**
1.1	[Matter is what makes up all objects or substances] *and contains both volume and mass.*
2.2	*The author of this passage* ... [defines matter as the stuff that all objects and substances in the universe are made of].
3.2	[Matter is identified as being present everywhere and in all substances].
5.1	[Matter is all the objects and substances] *that take up space* [around us].
CU106 W=4	**Matter has volume and mass.**
1.1	[Matter] *is what makes up all objects or substances and* [contains both volume and mass].
2.3	*All matter has the ability to be detected and measured as* [it takes up space defined as volume and contains a certain amount of material defined as mass].
4.3	[Matter is anything that has mass and takes up space (volume)].
5.2	*Matter can be detected and measured because* [it contains volume and mass].
CU142 W=1	**Energy is essential.**
2.20	*Although it can't be seen or touched like matter,* [energy is very essential].
CU145 W=1	**Matter can be a solid, liquid or gas.**
4.5	[Common examples are solids like a bed and a bug, liquids like water and milk, and gas like air and smoke].

TARGET: First Four Sentences of a Student Summary

		(1) In here universe, we use the matters to live with.
CU105	W=4	(2) [The matter is the stuff of all objects in the universe.]
CU106	W=4	(3) [Because every matter has volume. (4) And has mass,] matter be detected and measured.

Fig. 1 Four Content Units (CUs) with the CU id, weight and label in bold, and the first three sentences of a target student summary with matches to two of the illustrated CUs. The contributors from the model summaries, indexed in column 1 by the summary id (first field) and sentence id (second field), are in brackets. For clarity, we show the full sentence contexts, with the non-contributing part of the sentences in italics. The second and third sentences of the target student summary match CU105 and CU106, respectively. Note that the complete model had sixty CUs: three of weight 5, seven of weight 4, thirteen of weight 3, fifteen of weight 2, and twenty-two of weight 1

is not part of the CU. A contributor can be discontinuous, as in contributor 5.1 of CU105.

The content model is used to assess the content in new or *target* summaries. The same student who created the model exemplified in Fig. 1 annotated the twenty target summaries investigated in (Passonneau et al. 2013). To assess the content of a target summary against the model, the annotator selects phrases from the target that express the same meaning as CUs in the model. The raw score of a summary is the sum of the weights of all the matched CUs. For example, sentence 2, and sentence 3 plus part of 4 of the Student Summary (TARGET) in Fig. 1 each match a CU of weight 4; together they contribute 8 to the total raw score. After matching phrases from the target summary to model CUs, any text that fails to match CUs is then segmented into clauses, each corresponding to a CU of weight 0. For example, sentence 1 of the target in Fig. 1 adds to the CU count, but not to the summed weights.

For purposes of illustration, assume we have a target summary with fifteen CUs, and that the sum of the CU weights is 40. This raw score has two normalizations that measure quality and coverage. The quality score normalizes the raw score by the maximum sum that could be assigned to an equal number of CUs as in the target,

using each CU weight no more than w_j times, where w_j is the total number of model CUs of weight j. It indicates how close the target sum is to the maximum sum a student could achieve, given the same number of CUs. Given the distribution of CU weights listed in the caption of Fig. 1, the maximum sum for $N = 15$ would consume all three CUs of weight 5, all seven of weight 4, and five of weight 3, which sums to 58 $((3 \times 5) + (7 \times 4) + (5 \times 3) = 58)$. This gives a quality score of $\frac{40}{58} = 68.97\ \%$. The coverage score normalizes the raw sum by the maximum sum for the average number of CUs in a model summary. It indicates how close the raw score of the target is to that of an average wise-crowd model. Here the model summaries had an average of 27 CUs, with a maximum sum of 90, giving 44.44 % as the coverage score for our hypothetical summary. We also report the average of the content quality and coverage scores as a single comprehensive score, which here would be 56.71 %.

Concurrent with development of the wise-crowd method, we conducted multiple tests to determine how many model summaries would yield stable scores. All yielded the same result of between four and five models (Nenkova et al. 2007). For three different summarization tasks, content models were created using nine model summaries each, and scores based on these models were taken as reference scores. From 45 summaries on each of the three tasks, 68 pairs had a difference in reference scores of ≥ 0.1 (i.e., given nine model summaries). We re-computed scores for each member of these divergent pairs, varying the number of model summaries from one to eight, using all combinations of model summaries for each value of $n \in [1, 8]$. In the first test, we computed the average score for each summary given all content models with n model summaries, and its standard deviation. We determined that for each divergent pair, at least five model summaries were required for the confidence intervals of their average scores to diverge. In the second test, we investigated the probabilities of the three types of ranking errors: finding scores to be essentially equivalent when the reference scores differ, the converse, and misranking. The conditional probabilities of each type of error dropped to 0.05 or below at five models. Finally, correlation of scores with scores from nine models surpassed 0.80 at four models. From these three tests, we concluded that four to five model summaries were sufficient to produce stable scores.

Wise-crowd content assessment has been found to have high interannotator reliability of the two forms of annotation: creation of the model and annotation of target summaries.[3] Regarding the content model, reliability of models from two annotators working independently was measured using the agreement coefficient Krippendorff's alpha (Krippendorff 1980), combined with a distance metric (MASI) that gives partial credit to overlapping sets (Passonneau 2006). To measure reliability, we consider each CU, without the descriptive label, as a set of words selected from the model summaries by the annotator. We ask to what degree different annotators create the same disjoint sets. For each word in the model summaries, its annotation value is the set of words it is assigned to, less the word itself. Consider the contributor to CU142 in Fig. 1 (the last four words of the 20th sentence in model summary two): *energy is*

[3]It also produces consistent rankings of summarization systems given different sets of annotations, which is less relevant here (Passonneau 2010).

very essential. The annotation value of the word instance *energy* consists of *is very essential*. If a second annotator creates an identical set for this token of *energy*, the two annotators agree on the annotation value for this word. In Passonneau (2010), alpha_{MASI} ranged from 0.68 to 0.89 for ten pairs of models generated by different annotators. Cicchetti (1994) considers values above 0.60 to be good, and above 0.75 to be excellent; for Krippendorff (1980), the corresponding thresholds are 0.67 and 0.80. On the *What is Matter* task, reliability was 0.87 for content models from two student annotators who worked independently. To assess target annotation, we found an average alpha of 0.78 for three distinct summarization tasks from sixteen systems (Passonneau 2010). Finally, scores assigned to system summaries based on models and target annotations from distinct annotators on eight summarization tasks had a high average correlation of 0.86 (macro-level reliability).

Automated Wise-Crowd Content Assessment

We present two automated methods for wise-crowd content assessment, first reported in pilot studies and developed further here. Both methods address: 1) identification of sub-sentence units in summaries; 2) representation of their meaning; 3) measurement of meaning similarity; 4) selection of the best assignment of sub-sentence units to CUs from many possibilities. PyrScore automates the assessment of target summaries, given an existing content model (Passonneau et al. 2013). It segments sentences into word sequences, and represents meaning using a distributional semantic method developed for sentence similarity. PEAK constructs a content model and assesses targets using an analytic approach (Yang et al. 2016). It analyzes sentences into triples that represent subject-predicate-object relations, using an existing resource. To represent meaning and measure similarity, it uses a tool that consults word senses in WordNet (Fellbaum 1998), an electronic lexicon with rich semantics. PEAK compares the meanings of triples, or individual elements of triples, using a hypergraph data structure to keep track of the elements of the same triple. Optimization algorithms are used to select the optimal sentence segmentations for PyrScore, or optimal sets of triples for PEAK. Parameters for each method were tuned on the development set of twenty summaries from (Passonneau et al. 2013; Yang et al. 2016).

PyrScore

PyrScore identifies subsentence units by segmenting sentences into a covering set of word sequences, or ngrams. This simple method is easy to implement, and was found to work well (Passonneau et al. 2013). For computational efficiency, we restrict the maximum ngram length N to be no more than 14, which we found to be sufficient on the development set of twenty target summaries. The top of Fig. 2 schematically represents the segmentation of sentence 2 from the target summary illustrated in Fig. 1 into all possible ngram sequences.

In our previous work, we experimented with three semantic similarity methods (Passonneau et al. 2013). A word overlap score and a string edit distance compared

SENTENCE (11 words)
The matter is the stuff of all objects in the universe.

SEGMENTATIONS
One of: 11 unigrams
`the | matter | is | the | stuff | of | all | objects | in | the | universe`
Ten of: 1 bigram, 10 unigrams
`the matter | is | the | stuff | of | all | objects | in | the | universe`
`the | matter is | the | stuff | of | all | objects | in | the | universe`
...
One of: one 11-gram
`the matter is the stuff of all objects in the universe`

(SEGMENTATION, CU SET) PAIRS
the matter is the stuff | of all objects in the universe
{ 144, 1, MID }, { 105, 4, HIGH }
the matter | is | the stuff of all objects in the universe
{ 106, 4, LOW }, { }, { 103, 4, HIGH }

Fig. 2 First two stages of PyrScore: 1. SEGMENTATIONS: All segmentations for each sentence are computed, up to a threshold length ngram, which here is 14. 2. (SEGMENTATION, CU SET) pairs: Each substring in a segmentation is associated with a CU represented by a three-element set: the CU id, its weight, and the median cosine similarity of the vector representations of the substring and CU, where the cosine similarity exceeds a threshold

the actual words in an ngram with words in CUs. Here we rely exclusively on the third LSA-like method, Weighted Matrix Factorization (WMF) (Guo and Diab 2012), which had the best performance. Many low-dimensional methods such as LSA perform less well on similarity of sentences or phrases than on paragraphs, because of the limited number of words in the contexts, and the resulting sparsity of the vector representations. Unlike LSA, which effectively allows missing and observed words to have an equal impact on the matrix reduction, WMF assigns a small positive weight to missing words. The latent semantics has a larger context, but still depends more heavily on the observed words. WMF performs well on sentence similarity with a 100-dimension vector (Guo and Diab 2012). While much work uses 300-dimensions or more (Mikolov et al. 2013), much shorter latent semantic vectors have done well for tasks like Named Entity Recognition (Ma and Hovy 2016). We prepared a domain-independent corpus balanced across topics and genres, drawn from WordNet sense definitions, Wiktionary sense definitions, and the Brown corpus. From the corpus, a co-occurrence matrix M is constructed of unique words by sentences of size $I \times J$. Each cell M_{ij} holds a weighted count of word w_i in sentence s_j. We rely on the most commonly used weighted count, tf*idf (term frequency times inverse document frequency; idf discounts words that occur in all documents). Through matrix reduction, a 100-dimension latent vector representation is learned for every contributor and CU label in a content model, and for every ngram from target summaries (up to length 14; see above).

Because a CU contains multiple phrases to compare each ngram with, semantic similarity (SS) of an ngram n to a CU c is computed as an aggregate function $SS = G(n, c)$. As a result, different ways of expressing the same content are more likely to have high similarity. First, all the pairwise cosine similarities are computed

between WMF vectors of an ngram with the WMF vectors of the CU contributors and labels. Because we include the label, for a given weight w there will be $w + 1$ pairwise similarities to aggregate. PyrScore can be configured for G to be the maximum, minimum, mean or median of the set of similarity scores between the target substring and the elements of a CU. Comparisons where SS is greater than a threshold t are retained as candidates for the final solution. Parameters such as t that are tuned on the development set are summarized at the end of this subsection.

Figure 2 shows two segmentations of the target sentence from Fig. 1, and the set of CUs assigned to each segmentation; see the (SEGMENTATION, CU SET) PAIRS. The CU information includes the CU id, its weight, and SS value. To simplify the illustration, we illustrate two possible segmentations, using the values HIGH, MID, LOW instead of actual numeric values. The empty set is assigned to an ngram (e.g., *is*) if there is no CU where its SS exceeds the threshold t. The (segmentation,CU set) pairs are ranked by a score that is a function of the average length of the ngrams in the segmentation, the average weight of the CUs, and the average semantic similarities. The 6gram *of all the objects in the universe* matches CU105 from Fig. 1, which has weight 4, with SS=HIGH. Here, the first segmentation would have a higher average SS ($\frac{MID+HIGH}{2} > \frac{MID+LOW}{2}$) and higher average length ($5.5 > 3.7$). It gets ranked higher even though the second segmentation has a somewhat higher average CU weight ($2.7 > 2.5$) (note that a segment with no CU lowers the average CU weight). Only the top K such pairs are considered. Here, the first segmentation with two CUs was assigned to the sentence, whereas the human annotator assigned CU105 to the entire sentence. PyrScore occasionally posits extra ngram-CU matches in this way, which presumably results from the optimization objective to maximize the raw score.

The previous version of PyrScore used dynamic programming to select an optimal solution from candidate sentence segmentations and the corresponding CUs (Passonneau et al. 2013). We observed, however, that the allocation step is equivalent to a set cover or packing problem. In a set cover problem, the task is to cover a universe U of entities with a collection of subsets that obey a constraint, such as to use the smallest number of subsets whose union is U. In a packing problem, the task is to distribute a set of objects (e.g., CUs) into containers (e.g., sentences). We took advantage of the large literature on algorithmic solutions to these problems to arrive at a more general and efficient implementation. For example, the method we chose allocates CUs to sentences under the constraint that no CU can be used twice for one summary. Often, the same CU will have high cosine similarity with substrings from different sentences. In the previous implementation, this constraint was handled by post-processing.

The set cover and packing problems, each of which has many variants, are NP-complete. For an approximate solution, we rely on WMIN, a greedy algorithm that applies to a weighted graph in which each segmentation is a node, and one of the K segmentations per sentence is selected so as to output a maximal independent set (Sakai et al. 2003). On the development set, we experimented with three node weightings: the sum of the weights of the CUs associated with each node, the average cosine similarities, and the maximum cosine similarity. The edges in the WMIN graph in Fig. 3 represented by solid lines connect the top K (segmentation,CU set) pairs; at most one can be selected per sentence. The dashed edges connect nodes whose

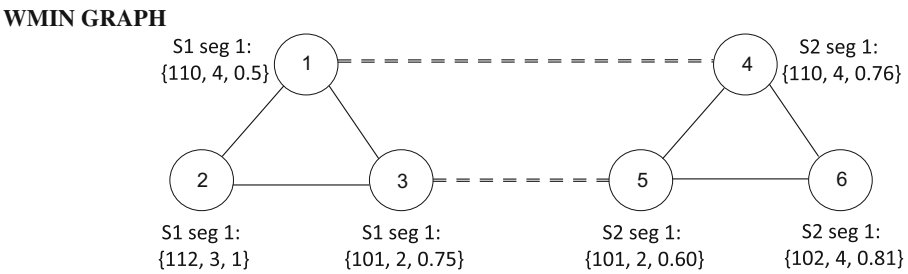


Fig. 3 Third stage of PyrScore: WMIN GRAPH: The nodes in the graph represent the top 3 segmentations for each sentence; edges connect all the segmentations for a sentence (solid lines), and all nodes that have overlapping CUs (dashed lines). For simplicity, the CU sets shown here have only one member CU. The pairs of CUs this graph allows are: (1,5), (1,6), (3,4), (3,6), (2,4), (2,5), (2,6)

CU sets have CUs in common, where only one can appear in a given solution. In this graph, the possible pairs of nodes for sentences one and two are: (1,5), (1,6), (2,4), (2,5), (2,6), (3,4), (3,6). Their ranking would depend on the node weights.

PyrScore has five parameters: 1. how many (segmentation, CU) pairs are considered for input to WMIN (K), 2. the minimimum ngram length of a segment (min), 3. the cosine similarity threshold for a CU match (c), 4. whether to use the minimum, maximum or median of the similarity scores between a segment and the elements of a CU (match), and 5. how the nodes in the WMIN graph are weighted (nodeW = sum of weights, or max cosine, or average cosine). We tested a large number of configurations and found the best correlations with the manual scores for ($K = 5$, min= 2, $c = 0.55$, match=median, nodeW = sum). Other configurations, however, can give similar results, as discussed below.

PEAK

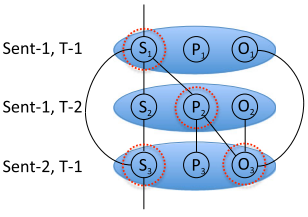
PEAK (for Pyramid Evaluation via Automated Knowledge Extraction) constructs a wise-crowd content model from model summaries, and uses it to assess new summaries (Yang et al. 2016). For both tasks, ClauseIE (Corro and Gemulla 2013) is used to automatically extract *<subject, predicate, object>* (s,p,o) triples; other OpenIE tools could be substituted. To measure similarity between individual elements from different triples, or between complete triples, it uses Align, Disambiguate and Walk (ADW), a tool that computes similarities of words, phrases, sentences or documents (Pilehvar et al. 2013). Like WMF, ADW performs well on sentences and subsentence units. A key advantage is that it requires no training. All parameters are those used by Yang et al. (2016), which include semantic similarity thresholds, and degree centrality of nodes in a graph data structure used to organize triples and their similarities to each other.

Content Model Generation To construct a content model, ClauseIE is first applied to all sentences in the model summaries, and all extracted triples become hyperedges in a hypergraph. A hypergraph is a generalization of a graph in which hyperedges can connect any number of nodes. The nodes in PEAK hyperedges are the individual

elements of (s,p,o) triples; ClauseIE sometimes extracts (s,o) pairs instead of triples. Distinct triples can share up to two nodes, so hyperedges can overlap. For example, two triples are generated from the model summary sentence, *[[Energy]s_1]s_2 [[can]p_1]p_2 only [[be transferred from one object or system]o_1 to another]o_2*, as indicated by the indexed brackets.

The hypergraph is extended with ordinary edges that represent ADW similarity (S_{adw}) between pairs of vertices v. $S_{adw}(v_i, v_j) \in [0, 1]$ is computed for all pairs of nodes not in the same hyperedge and a similarity edge (v_i, v_j) is added where $S_{adw}(v_i, v_j) \geq M_1$, for $M_1 = 0.5$, meaning the nodes are more similar than not. The left image in Fig. 4 schematically represents a PEAK hypergraph, with shaded elipses representing the hyperedges, and solid lines for S_{adw} edges. Sentence one (Sent-1) yields two triples whose nodes are all distinct, as indicated by the distinct indexing of the s, p, o nodes. To find triples that are more likely to be similar to multiple distinct triples, and thus generate CUs, we define nodes in the hypergraph to be relatively more salient if the similarity degree (number of similarity edges) is greater than d. Here we use $d = 3$, the midpoint of possible CU weights; with a larger number of model summaries, d would be increased. The nodes circled in dashed lines represent salient nodes.

Every triple in the hypergraph that has at least two salient nodes (e.g., Sent-2, T-1) serves as an anchor t_i for a candidate CU. For each anchor t_i from a model summary S_x, triples from other model summaries are evaluated as possible contributors. We define the *similarity class* for each node of an anchor to be all the nodes it is connected to by a similarity edge. If a triple t_j from a distinct model summary S_y has two or more nodes in the similarity classes of t_i, the overall similarity of t_j to t_i is the sum of the nodes' similarity scores. The t_j with the highest summed similarity is selected as S_y's contributor to the CU anchored by t_i (from S_x). Finally, the candidate CU is represented as the anchor triple t_i along with the number of contributors as its weight, omitting the contributors. Each triple t_j that contributes to t_i's CU is in turn a potential anchor, so for a candidate CU with n contributors there can be up to n variants. These are merged into a single CU that includes all the contributors such that for any pair t_i, t_n, a pair of nodes v_i in t_i and v_n in t_n have an ADW similarity greater than a threshold M_3. The final weight is the maximum weight associated with any of



	s	p	o
Anchor	matter	is	all the objects and substances
Contributor 1	matter	is	all the objects and substances
Contributor 2	matter	is identified	as being present everywhere
Contributor 3	matter	as the stuff	in all objects and substances
Contributor 4	matter	makes up	all objects or substances

Fig. 4 The hypergraph on the left contains (s,p,o) hyperedges (shaded ellipses), and s,p,o nodes from the same triple are co-indexed. Solid lines connect pairs of nodes from distinct hyperedges that are semantically similar. Dashed circles identify the salient nodes whose semantic similarity degree is ≥ 3. On the right is a table showing a CU generated by PEAK, prior to the merging step. Each contributor is a triple that had at least two nodes whose similarity to nodes in the anchor was ≥ 0.5

the anchors. We use $M_3 = 0.8$ based on the pilot experiment with the development set (Yang et al. 2016).

Scoring Target Summaries To score target summaries, $< s, p, o >$ triples are extracted from each target summary. The semantic similarity of target triples to model CUs is computed using ADW, after the lexical content of the (s,p,o) nodes in each model or target triple is concatenated into a single string. To find the maximal score for a target summary while ensuring that the assignment between target triples and CUs is one-to-one is a maximal matching problem. To find a solution, we apply the Munkres-Kuhn algorithm, also known as the Hungarian algorithm (Kuhn 1955). A bipartite graph is constructed from CU nodes u to nodes v that represent target summary triples. An edge (u, v) exists if $S_{adw}(u, v) \geq T$. We use $T = 0.6$, again based on Yang et al. (2016). The cost of selecting (u, v) is the weight of the CU u.

Comparison of Main Ideas and Wise-Crowd Assessments

For the *What is Matter?* task, we have 120 new summaries from the Perin et al. (2013) study to compare the main ideas rubric with wise-crowd content assessment. A random subset of twenty used earlier (Passonneau et al. 2013; Yang et al. 2016) are the current development set (Devel. 20), and are excluded from the evaluation set of 120 (Test 120). The manual wise-crowd assessment was applied to a new random subset of twenty (Test 20). We pose the following questions:

1. Does the manual wise-crowd content assessment correlate well with the main ideas score?
2. Do the automated wise-crowd methods correlate with the manual method?
3. Do the automated wise-crowd methods correlate with the main ideas score?
4. How do the automated methods compare?

For questions 1 and 2, we use the subset of twenty new summaries that were manually scored against the wise-crowd content model (Test 20). For questions 3 and 4 we use Test 120. Again, we use the same wise-crowd content model from Passonneau et al. (2013).

Question 1: Main Ideas Rubric versus Manual Wise-Crowd Content Assessment

All fourteen ideas from the main ideas rubric occur in the CU model with weights ≥ 3, but not one-to-one: they match the three weight 5 CUs, four out of seven of the weight 4 CUs, and six of the thirteen weight 3 CUs (see Fig. 1). This semantic overlap makes it unsurprising to find a high correlation between the rubric and the manual wise-crowd content assessment. On the twenty new randomly selected summaries, the Pearson correlation of the raw wise-crowd score with the main ideas score is 0.88, as in Table 1; the first author did the annotation. For the development set of the twenty original summaries used in (Passonneau et al. 2013), the correlation is 0.90. The correlations of main ideas with the three normalized wise-crowd scores (see

Table 1 Pearson correlations of the main ideas rubric with manual wise-crowd content scores for 20 pilot summaries and 20 development summaries

	Devel. 20	Test 20
Manual raw	0.90	0.88
Manual quality	0.79	0.81
Manual coverage	*0.90*	*0.88*
Manual comprehensive	0.89	0.88

Note: the correlations for the pairs of raw scores and pairs of coverage scores are necessarily the same, so the latter appear in italics

above) are also shown; the quality score correlation is somewhat lower. Note that the normalization for content coverage uses the same normalization factor for all target summaries; therefore, the content coverage score has the same correlation as the raw score. The content quality score, however, uses a distinct normalization factor for each target summary (see above).[4]

Question 2: Manual versus Automated Wise-Crowd Content Assessment

As shown in Table 2, the four PyrScore scores (raw, quality, coverage, comprehensive) correlate very well with the manual wise-crowd content scores on the test set of twenty summaries, especially the raw score (Pearson=0.95). The correlations with the development set of twenty are also shown.

In contrast to PyrScore, PEAK is fully automated: it generates a content assessment model, and then assesses writing samples relative to this model. It produces 82 CUs instead of 60, with many semantic variants of the same CU. For example, it has eighteen weight 5 CUs in comparison to three in the manual content model. On the Test 20, the correlation of the PEAK raw score with the manual raw content score is 0.82, a very positive outcome for a fully automated method. We modified PEAK to produce quality, coverage and comprehensive content scores; the quality and comprehensive scores correlate less well.[5]

The mean raw score for manual assessment is 27.65 compared with 36.55 for PyrScore and 31.55 for PEAK. Two-tailed paired t-tests to compare the means indicate that the PyrScore mean is significantly different (t=-7.76, p=2.6×10^{-7}), while the difference in the PEAK mean nearly reaches significance (t=-1.90, p=0.07); the same pattern occurs on Devel. 20. While this is too small a sample for strong conclusions, it seems clear that the accuracy of the automated methods could be improved.

[4]The correlation of the comprehensive score from Passonneau et al. (2013) was 0.85, which has been corrected to 0.89.

[5]The PEAK results reported in Yang et al. (2016) rely on an earlier version of ADW. The score correlation of 0.81 for Devel. 20 reported there is higher than the 0.78 we get here.

Table 2 Pearson correlations of the two automated methods, PyrScore and Peak, with the manual wise-crowd content assessment

	PyrScore		PEAK	
	Devel. 20	Test 20	Devel. 20	Test 20
Manual raw	0.84	0.95	0.78	0.82
Manual quality	0.74	0.81	0.52	0.47
Manual coverage	*0.84*	*0.95*	*0.78*	*0.82*
Manual comprehensive	0.85	0.90	0.55	0.46

Question 3: Main Ideas Rubric versus Automated Wise-Crowd Content Assessment

Given the high correlation of main ideas and manual wise-crowd content assessment, and of the manual and automated wise-crowd content assessment, we expect good correlation of main ideas with the automated methods as well. Table 3 shows the correlations of the scores from the automated methods with the main ideas score for the 120 summaries in the test set. The quality score has poor correlations, mainly because the quality normalization depends on an accurate count of total CUs in the target summary, and neither method was engineered to do this well. For the other scores, PyrScore correlates best, at 0.83 for the raw score and 0.82 for the comprehensive score. PEAK has a relatively lower but still reasonable correlation of 0.70 for the raw score.

Question 4: Comparison of PyrScore and PEAK

The raw scores of both automated methods correlate well with the manual wise-crowd content assessment and with the main ideas score. PyrScore has the higher correlations in both cases, which is unsurprising in that it relies on the manual content model. As we saw above, the manual content model contains all fourteen of the main ideas in the main ideas rubric. All but one also occur in the content model that PEAK constructs automatically. PEAK assigns weights to these thirteen "main ideas" CUs that are the same as the manual content model in seven cases, one higher in four cases, and one lower in two cases.

Another way to compare the automated and manual wise-crowd scores is by their absolute values. Both methods yield higher absolute values than the manual method

Table 3 Pearson correlations between 120 main ideas scores and automated wise-crowd scores

	Test 120	
	PyrScore	PEAK
Raw	0.83	0.70
Quality	0.57	−0.17
Coverage	*0.83*	*0.70*
Comprehensive	0.82	0.24

with statistical significance, or approaching statistical significance. Interestingly, the PEAK comprehensive score, which correlates less well with the main ideas score, has the same mean ($\mu = 0.44$).

PyrScore results can be compared directly with the manual scoring to judge how often annotators and PyrScore assign the same CUs to target summaries. PyrScore has a moderately good average recall of 0.62 CUs, where recall is the proportion of manually assigned CUs that PyrScore assigns. It has poor average precision of 0.45, where precision is the proportion of assigned CUs that are correct. The poor precision is consistent with our earlier observation that PyrScore assigns spurious extra CUs. F-measure, the harmonic mean of recall and precision, is 0.55. On the development set, other high-performing configurations of PyrScore have identical values of *c* and *match*, and otherwise similar values, e.g., where $K \in [3, 4]$, min=3, nodeW $\in$ {max, avg}. The three next best configurations have worse recall, but much better precision and better F-measure. To investigate this comparison of competing configurations further, we computed weighted recall and precision, multiplying each increment to the correct or incorrect CU count by the CU weight. The configuration we use here has much better average weighted recall than the three closest competing configurations, and the same F-measure. This suggests that the reason the selected configuration has better score correlations is its higher recall for high-weighted CUs.

PyrScore cannot be used with a PEAK content model, and PEAK target assessment cannot be done with a manual content model, because the CU representations are not commensurable. A key difference is that PyrScore computes pairwise semantic similarity of a target ngram with each element in a CU, but PEAK CUs have only one element. For PyrScore to rely on comparison to a single CU element, or for PEAK to rely on comparision to multiple CU elements, would require non-trivial changes to the assignment algorithms, and would presumably be disadvantageous. Another key difference is that PyrScore considers all possible ngram segmentations of a target summary sentence, whereas PEAK extracts (s,p,o) triples. It is possible, however, that a hybrid method might perform better than either does alone.

Discussion

The experimental results presented in the preceding section show that wise-crowd content assessment applied manually has a high Pearson correlation (0.88) with a rubric to assess whether students' summaries articulate the main ideas of a science passage. The raw scores from the two automated methods, PyrScore and PEAK, also correlate well with the main ideas score (0.83 and 0.70, respectively). Instead of deriving a content rubric from source texts, wise-crowd content assessment relies on a set of model writing samples. We believe this can be justified as follows. The purpose of the community college curriculum was to compensate for students' lack of preparation for postsecondary education. Because the wise-crowd members were masters students at a highly competitive graduate school, the resulting content model represents a level the target population aims to achieve. The high correlation with a rubric that was successfully used to measure the positive impact of a particular

instructional method indicates that wise-crowd content assessment could replace the rubric.

Comparison of PyrScore results and manual annotation indicate it has moderately good recall (0.62) but low precision (0.45) for specific CUs. To give teachers or students feedback on the individual CUs that their summaries express or omit would require improvements to the automated methods. However, automated wise-crowd content assessment could potentially provide useful feedback in educational settings, based on the distinction between content quality and content coverage. If the coverage is low but the quality is high, a student could be informed that the ideas expressed in the summary are good, but more needs to be said. The converse would be true if the student has included all the important ideas, but in addition, too many unimportant ones. In future work, we plan to investigate whether teachers and students can benefit from such feedback.

There are near-term and long-term advantages to substitution of wise-crowd content assessment for a content rubric, or as part of a more complex rubric, such as those discussed in the section, *Summarization Rubrics for Educational Interventions*. Researchers or teachers would not have to develop and test a new content rubric, which could facilitate more numerous and extensive studies of writing skills. In addition, through the use of a common method across studies, results across studies would be more comparable, which would support generalization of conclusions. We discuss the preconditions for current usage in the next paragraph. In the long term, there is the possibility of integration of wise-crowd content assessment with online learning environments to provide feedback for teachers and students. Such feedback, however, would depend on redesigning the software to produce results in real time. To process 120 summaries on a laptop with 4 core 2.6GHz i5-3230M CPU, PyrScore took about thirty minutes. PEAK took closer to a day on a somewhat less powerful laptop, mainly due to the processing time for ADW. Optimization of both software packages is under development.

Figure 5 tabulates what is needed to apply the methods. For a given set of writing samples, the main precondition is to collect writing samples from a wise crowd of at least five members. The members of the wise crowd can be selected according to criteria established by the researcher. Their samples should be written in response to the same prompt or stimulus, and they should follow the same instructions, as the population to be assessed. The annotation tool to create a manual model generates raw and quality scores, but not the coverage score. For manual application of the method, at least two annotators should be trained in the guidelines and use of the

	Manual	PyrScore	PEAK
Prompt/Instructions	y	y	y
Wise Crowd Samples	y	y	y
Create Manual Model	y	y	n
Raw Score	y	y	y
Quality Score	y	y	y*
Coverage Score	n	y	y*
Comprehensive Score	n	y	y*
CU accuracy	y	y	y
Matched CUs	y	y	n

Fig. 5 Preconditions and outputs of wise-crowd content assessment. Note that a new version of PEAK produces the scores indicated by y* that are not yet part of the download package

annotation tool to assess inter-annotator reliability. A downloadable package that can run PyrScore has been completed, and will be made available. PEAK is already available for download, and is being upgraded to make it more user-friendly.[6]

Conclusion

Although automated methods to assess student writing have been around since the work of Page (1966) and Page (1968), relatively little of this work has addressed how to automatically detect specific ideas. Apart from the handful of references cited here, there has been little effort to incorporate automated methods into instructional methods and environments for writing skills. We have demonstrated how wise-crowd content assessment specifies a content model for a given writing prompt. A distinctive aspect of the representation is that units of content are differentiated by their weights, which captures a critical aspect of text comprehension: that propositions expressed in text are not equally important. The automatic techniques to apply this content assessment have been shown to correlate well with the manual content assessment, and also with a completely independent rubric to assess the main ideas of students' essays.

A key aspect of the approach is that it assesses content relative to an emergent model based on appropriate selection of a wise crowd, rather than on an externally specified standard. Thus it establishes a standard through exemplars that must be hand-selected, rather than on criteria the examples should meet. This could be viewed as a limitation for cases where it is preferable for experts to establish specific criteria for the content. This limitation can partly be addressed, however, by having experts produce the wise-crowd writing samples in a fashion that meets their criteria. The resulting model yields an explicit representation of the desired content, as well as different expressions of that content.

Compared with previous work that tests automatic methods to apply content assessment to students' writing (Gerard et al. 2016; Liu et al. 2014; Beigman-Klebanov et al. 2014; Rosé and VanLehn 2005; Butcher and Kintsch 2001), our work has the following advantages. First, it requires far less training data: on the order of a half dozen samples rather than hundreds or thousands. Second, the wise-crowd automated methods have high correlations with the manual methods, and both have high correlations with a rubric used in a writing intervention study. Finally, it produces two content scores that provide detailed information about quality and coverage of specific ideas.

Acknowledgments This paper is an extended version of an oral presentation made at an NSF-funded workshop held May 7-8, 2015 entitled MARWiSE: Multidisciplinary Advances in Reading and Writing for Science Education (Award IIS-1455533). The authors thank members of the workshop for their constructive feedback. We also thank Weiwei Guo for input regarding his Weighted Matrix Factorization method, and his suggestions for related work. Finally, we thank three anonymous reviewers for their constructive criticism.

[6]Downloadable packages for PyrScore and PEAK will be available from the Columbia University Academic Commons, and The Pennsylvania State University Data Commons, where Passonneau has recently moved.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater®V. 2. *Journal of Technology Learning and Assessment*, *4*(3), 3–39.

Bangert-Drowns, R.L., Hurley, M.M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*(1), 29–58.

Beers, S.F., & Nagy, W.E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre? *Reading and Writing*, *22*, 185–200.

Beers, S.F., & Nagy, W.E. (2011). Writing development in four genres from grades three to seven: syntactic complexity and genre differentiation. *Reading and Writing*, *24*, 183–202.

Beigman-Klebanov, B. (2015). Towards automated evaluation of writing along STEM-relevant dimensions. MARWiSE: Multidisciplinary Advances in Reading and Writing for Science Education Workshop. May 7-8, 2015, Columbia University.

Beigman-Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 247–252*.

Berland, L.K., & McNeill, K.L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, *94*(5), 765–793.

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.

Brown, A.L., Day, J.D., & Jones, R.S. (1983). The development of plans for summarizing texts. *Child Development*, *54*, 968–979.

Burstein, J., Elliot, N., & Molloy, H. (2016). Informing automated writing evaluation using the lens of genre: Two studies. *CALICO Journal*, *33*(1).

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Enriching automated essay scoring using discourse marking, In Stede, M., Wanner, L., & Hovy, E. (Eds.) *Workshop on Discourse Relations and Discourse Marking, pages 15–21. Association for Computational Linguistics*.

Butcher, K.R., & Kintsch, W. (2001). Support of content and rhetorical processes of writing: Effects on the writing process and the written product. *Cognition and Instruction*, *19*(3), 277–322.

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychologial Assessment*, *6*(4), 284–290.

Corro, L.D., & Gemulla, R. (2013). ClausIE: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13), pages 355–366*.

Day, J.D. (1986). Teaching summarization skills: Influences of student ability level and strategy difficulty. *Cognition and Instruction*, *3*(3), 193–210.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). Cognitive models of writing: Writing proficiency as a complex integrated skill. Technical Report 2, ETS Research Report Series, Princeton, NJ.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA*.

Foltz, P.W., Laham, D., & Landauer, T.K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2).

Garner, R. (1985). Text summarization deficiencies among older students: Awareness or production ability *American Educational Research Journal*, *22*(4), 549–560.

Gerard, L.F., Ryoo, K., McElhaney, K.W., Liu, O.L., Rafferty, A.N., & Linn, M.C. (2016). Automated guidance for student inquiry. *Journal of Educational Psychology*, *108*(1), 60–81.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H.I. (2010). Understanding and integrating multiple science texts: Summary tasks are sometimes better than argument tasks. *Reading Psychology*, *31*(1), 30–68.

Gillespie, A., Graham, S., Kiuhara, S., & Hebert, M. (2014). High school teachers' use of writing to support students' learning: a national survey. *Reading and Writing*, *27*(6), 1043–1072.

Glymph, A. (2010). *The nation's report card: Reading 2009. Technical Report NCES 2010-458*. Washington: National Center for Education Statistics (NCES).

Glymph, A. (2013). *The nation's report card: Reading 2012. Technical Report NCES 2012-457*. Washington: National Center for Education Statistics (NCES).

Glymph, A., & Burg, S. (2013). *The nation's report card: A first look: 2013 mathematics and reading. Technical Report NCES 2014-451*. Washington: National Center for Education Statistics (NCES).

Graham, S., Capizzi, A., Harris, K., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: a national survey. *Reading and Writing*, *27*(6), 1015–1042.

Graham, S., & Perin, D. (2007a). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476.

Graham, S., & Perin, D. (2007b). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. New York: Technical report, Carnegie Corporation of New York.

Guo, W., & Diab, M. (2012). Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), pages 864–872*.

Hand, B.M., Hohenshell, L., & Prain, V. (2004). Exploring students' responses to conceptual questions when engaged with planned writing experiences: A study with year 10 science students. *Journal of Research in Science Teaching*, *41*(2), 186–210.

Hughes, S., Hastings, P., Magliano, J., Goldman, S., & Lawless, K. (2012). Automated approaches for detecting integration in student essays. In *Automated approaches for detecting integration in student essays. Springer-Verlag*.

Johnson, R.E. (1970). Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior*, *9*(1), 12–20.

Kellogg, R.T. (2008). Training writing skills: a cognitive development perspective. *Journal of Writing Research*, *1*(1), 1–26.

Kharkwal, G., & Muresan, S. (2014). Surprisal as a predictor of essay quality. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pages 54–60*.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 364–394.

Klein, P.D., & Rose, M.A. (2010). Teaching argument and explanation to prepare junior students for writing to learn. *Reading Research Quarterly*, *45*(4), 433–461.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Beverly Hills: Sage Publications.

Kuhn, H.W. (1955). The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, *2*, 83–97.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*.

Liu, O.L., Brew, C., Blackmore, J., Gerard, L.F., Madhok, J., & Linn, M.C. (2014). Automated scoring of constructed-response science items: Propsects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28.

Louis, A., & Nenkova, A. (2009). Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 306–314, Singapore. Association for Computational Linguistics*.

Louis, A., & Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, *39*(2), 267–300.

Ma, X., & Hovy, E.H. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). To Appear*.

Magliano, J.P., Trabasso, T., & Graesser, A.C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, *91*, 615–629.

Mazzeo, C., Rab, S.Y., & Alssid, J.L. (2003). *Building bridges to college and careers: Contextualized basic skills programs at community colleges. Technical report*. Brooklyn: Workforce Strategy Center.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality.

NCES (2012). The nation's report card: Writing 2011.

Nenkova, A., & Passonneau, R.J. (2004). Evaluating content selection in summarization: The pyramid method, In Susan Dumais, D.M., & Roukos, S. (Eds.) *HLT-NAACL 2004: Main Proceedings, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics*.

Nenkova, A., Passonneau, R.J., & McKeown, K. (2007). The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, *4*(2).

Norris, S.P., & Phillips, L.M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Olinghouse, N.G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology*, *107*(2), 391–406.

Olinghouse, N.G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, *26*(1), 45–65.

Owczarzak, K., Conroy, J.M., Dang, H.T., & Nenkova, A. (2012). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pages 1–9. Association for Computational Linguistics*.

Page, E.B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *47*(5), 238–243.

Page, E.B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*, 210–225.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of experimental education*, *62*(2), 127–142.

Passonneau, R., & Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, *2*, 311–326.

Passonneau, R.J. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC), pages 831–836, Genoa, Italy*.

Passonneau, R.J. (2010). Formal and functional assessment of the pyramid method for summary content evaluation. *Natural Language Engineering*, *16*, 107–131.

Passonneau, R.J., Baker, C.F., Fellbaum, C., & Ide, N. (2012). The MASC word sense corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12). European Language Resources Association (ELRA)*.

Passonneau, R.J., Chen, E., Guo, W., & Perin, D. (2013). Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 143–147, Sofia, Bulgaria. Association for Computational Linguistics*.

Passonneau, R.J., Goodkind, A., & Levy, E. (2007). Annotation of children's oral narrations: Modeling emergent narrative skills for computational applications. In *Proceedings of the Twentieth Annual Meeting of the Florida Artificial Intelligence Research Society (FLAIRS-20), pages 253–258*.

Passonneau, R.J., McKeown, K., & Sigelman, S. (2006). Applying the pyramid method in the 2006 Document Understanding Conference. In *Proceedings of the 2006 Workshop of the Document Understanding Conference (DUC)*.

Passonneau, R.J., Nenkova, A., McKeown, K., & Sigelman, S. (2005). Applying the pyramid method in DUC 2005. In *Proceedings of the 2005 Workshop of the Document Understanding Conference (DUC)*.

Perin, D., Bork, R.H., Peverly, S.T., & Mason, L.H. (2013). A contextualized curricular supplement for developmental reading and writing. *Journal of College Reading and Learning*, *43*(2), 8–38.

Persky, H.R., Daane, M.C., & Jin, Y. (2003). *The nation's report card: Writing 2002. Technical Report NCES 2003-529*. Washington: National Center for Education Statistics (NCES).

Pilehvar, M.T., Jurgens, D., & Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1341–1351, Sofia, Bulgaria. Association for Computational Linguistics*.

Proske, A., Narciss, S., & McNamara, D.S. (2012). Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading*, *35*(2), 136–152.

Qazvinian, V., & Radev, D.R. (2012). A computational analysis of collective discourse. In *Proceedings of the 2012 Conference on Collective Intelligence, Cambridge MA*.

Reiser, B.J., & Kenyon, L.K.B.L. (2012). Engaging students in the scientific practices of explanation and argumentation. *Science and Children*, *49*(8), 8–13.

Roscoe, R.D., Allen, L.K., Weston, J.L., Crossley, S.A., & McNamara, D.S. (2015a). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39–59.

Roscoe, R.D., & McNamara, D.S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*(4), 1–16.

Roscoe, R.D., Snow, E.L., Allen, L.K., & McNamara, D.S. (2015b). Automated detection of essay revising patterns: applications for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, and Learning*, *10*(1), 59–79.

Rosé, C., & VanLehn, K. (2005). An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. *International Journal of Artificial Intelligence in Education*, *15*(4), 325.

Rudner, L.M., Garcia, V., & Welch, C. (2006). An evaluation of Intellimetric^{TM} essay scoring system. *The Journal of Technology Learning and Assessment*, *4*(4).

Saggion, H., Torres-Moreno, J.-M., da Cunha, I., SanJuan, E., & Velázquez-Morales, P. (2010). Multilingual summarization evaluation without human models. In *Proceedings of Coling 2010, pages 1059–1067*.

Sakai, S., Togasaki, M., & Yamazaki, K. (2003). A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, *126*(2-3), 313–322.

Salahu-Din, D., Persky, H.R., & Miller, J. (2008). *The nation's report card: Writing 2007. Technical Report NCES 2008-468*. Washington: National Center for Education Statistics (NCES).

Sampson, V., Enderle, P., Grooms, J., & Witte, S. (2013). Writing to learn by learning to write during the school science laboratory: Helping middle and high school students develop argumentative writing skills as they learn core ideas. *Science Education*, *97*(5), 643–670.

Shermis, M., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and Future Directions*. New York: New York: Routledge.

Slotta, J.D., & Linn, M.C. (2009). *WISE Science*. New York: Teachers College Press.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Doubleday.

Teufel, S., & van Halteren, H. (2004). Evaluating information content by factoid analysis: Human annotation and stability, In Lin, D., & Wu, D. (Eds.) *Proceedings of EMNLP 2004, pages 419–426, Barcelona, Spain. Association for Computational Linguistics*.

Turner, A.A. (1987). *The propositional analysis system. Technical Report 87-2, University of Colorado*. Boulder: Department of Psychology and Institute of Cognitive Science.

Turner, A.A., & Greene, E. (1978). *The construction and use of a propositional analysis system. Technical Report JSAS Catalog of Selected Documents in Psychology, no. 1713*. Washington, DC: American Psychological Association.

van Dijk, T.A., & Kintsch, W. (1977). Cognitive psychology and discourse: Recalling and summarizing stories, In Dressier, W.U. (Ed.) *Trends in text-linguistics, pages 61–80. De Gruyter, New York*.

van Dijk, T.A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

VanLehn, K., Jordan, P., & Rosé, C.P. (2002). The architecture of Why2-Atlas: a coach for qualitative physics essay writing, In Cerri, S.A., Gouarderes, G., & Paraguacu, F. (Eds.) *Intelligent Tutoring Systems, 2002: 6th International Conference, pages 158–167, Berlin. Springer*.

Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders*, *30*(4), 275–287.

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

Wiley, J., & Voss, J.F. (1996). The effects of playing historian on learning in history. *Applied Cognitive Psychology*, *10*, 63–72.

Yang, Q., Passonneau, R.J., & de Melo, G. (2016). PEAK: Pyramid evaluation via automated knowledge extraction. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016). AAAI Press*.

Yore, L.D., Hand, B.M., & Florence, M.K. (2004). Scientists' views of science, models of writing, and science writing practices. *Journal of Research in Science Teaching*, *41*(4), 338–369.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge: Addison-Wesley.