Psicología Educativa
ISSN: 1135-755X
ISSN: 2174-0526
Colegio Oficial de Psicólogos de Madrid

Martínez-Huertas, José Ángel; Jastrzebska, Olga; Mencu, Adrián; Moraleda, Jessica; Olmos, Ricardo; León, José Antonio

Analyzing Two Automatic Latent Semantic Analysis (LSA) Assessment Methods (Inbuilt Rubric vs. Golden Summary) in Summaries Extracted from Expository Texts

Psicología Educativa, vol. 24, no. 2, 2018, July-December, pp. 85-92
Colegio Oficial de Psicólogos de Madrid

DOI: https://doi.org/10.5093/psed2048a9

Psicología Educativa (2018) 24(2) 85-92

Psicología Educativa

http://journals.copmadrid.org/psed

Analyzing Two Automatic Latent Semantic Analysis (LSA) Assessment Methods (Inbuilt Rubric vs. Golden Summary) in Summaries Extracted from Expository Texts

José Ángel Martínez-Huertas, Olga Jastrzebska, Adrián Mencu, Jessica Moraleda, Ricardo Olmos, and José Antonio León

Universidad Autónoma de Madrid, Spain

ARTICLE INFO

Article history:
Received 16 November 2017
Accepted 4 December 2017
Available online 16 April 2018

Keywords:
LSA
Inbuilt rubric
Automatic essay scoring (AES)
Lexical descriptors
Summaries

A B S T R A C T

The purpose of this study was to compare two automatic assessment methods using Latent Semantic Analysis (LSA): a novel LSA assessment method (Inbuilt Rubric) and a traditional LSA method (Golden Summary). Two conditions were analyzed using the Inbuilt Rubric method: the number of lexical descriptors needed to better accommodate an expert rubric (few vs. many) and a weighting function to penalize off-topic contents included in the student summaries (weighted vs. non-weighted). One hundred and sixty-six students divided in two different samples (81 undergraduates and 85 High School students) took part in this study. Students summarized two expository texts that differed in complexity (complex/easy) and length (1,300/500 words). Results showed that the Inbuilt Rubric method simulates human assessment better than Golden summaries in all cases. The similarity with human assessment was higher for Inbuilt Rubric (r = .78 and r = .79) than for Golden Summary (r = .67 and r = .47) in both texts. Moreover, to accommodate an expert rubric into the Inbuilt Rubric method was better using few descriptors and the weighted function.

Análisis de dos métodos de evaluación automática de análisis semántico latente (LSA): un nuevo método LSA (*Inbuilt Rubric*) y un método LSA tradicional (*Golden Summary*) en resúmenes extraídos de textos expositivos

Palabras clave:
LSA
Inbuilt rubric
Automatic essay scoring (AES)
Descriptores léxicos
Resúmenes

R E S U M E N

El objetivo de este estudio es comparar dos métodos de evaluación automática del análisis semántico latente (LSA): Un nuevo método LSA (*Inbuilt Rubric*) y un método LSA tradicional (*Golden Summary*). Se analizaron dos condiciones del método *Inbuilt Rubric*: el número de descriptores léxicos que se utilizan para generar la rúbrica (pocos vs. muchos) y una corrección que penaliza el contenido irrelevante incluido en los resúmenes de los estudiantes (corregido vs. no corregido). Ciento sesenta y seis estudiantes divididos en dos muestras (81 estudiantes universitarios y 85 estudiantes de instituto) participaron en este estudio. Los estudiantes resumieron dos textos expositivos que tenían distinta complejidad (difícil/fácil) y longitud (1,300/500 palabras). Los resultados mostraron que el método *Inbuilt Rubric* imita las evaluaciones humanas mejor que *Golden Summary* en todos los casos. La similitud con las evaluaciones humanas fue más alta con Inbuilt Rubric (r = .78 and r = .79) que con *Golden Summary* (r = .67 and r = .47) en ambos textos. Además, la versión de *Inbuilt Rubric* con menor número de descriptores y con corrección es la que obtuvo mejores resultados.

Latent Semantic Analysis (LSA) is a theory and a method for extracting and representing the meaning of words using statistical computations applied to a large corpus of text (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). Traditionally, LSA applied Singular Value Decomposition (SVD) as the linear algebra method to compute the similarity between words and groups of words. Since its beginnings in the 90s, LSA has been applied as a computational representation of the semantic memory for human-generated essays using Automatic Essay Evaluation (AEE). In this way, Landauer et al. (2007) presented LSA as a way to conceptualize text content in terms of connections among the words and word sequences within the text (O'Reilly & Munakata, 2000).

In recent years, the number of papers mentioning LSA has increased greatly (see, for example, Visinescu & Evangelopoulos, 2014) and new applications have been developed due to the similarity between the LSA and the human cognition, especially in

Cite this article as: Martínez-Huertas, J. A., Jastrzebska, O., Mencu, A., Moraleda, J., Olmos, R., & León, J. A. (2018). Analyzing two automatic assessment LSA methods (Inbuilt Rubric vs. Golden Summary) in summaries extracted from expository texts. *Psicología Educativa, 24*, 85-92. https://doi.org/10.5093/psed2048a9

Funding: This study was supported by Grant PSI2013-47219-P from the Ministry of Economic and Competitive (MINECO) of Spain, and European Union.
Correspondence: josea.martinez@uam.es (J. A. Martínez-Huertas).

semantic memory (e.g., Günther, Dudschig, & Kaup, 2015). Some of these applications are: identifying current tendencies in research (Aryal, Gallivan, & Tao, 2015; Wendy, How, & Atoum, 2014; Xu et al., 2015), improving search engines (Borisov, Serdyukov, & de Rijke, 2016; Ryan, Kaltman, Mateas, & Wardrip-Fruin, 2015), and producing keywords (Pu, Jin, Wu, Han, & Xue, 2015). LSA has also been applied in clinical domains, as automatically diagnosing psychological disorders (Cohen, Blatter, & Patel, 2008; Jorge-Botana, Olmos, & León, 2009) or improving tests used to prevent future risk of neuropsychological illness such as dementia (Pakhomov & Hemmy, 2014). LSA has been used in the linguistics and educational areas, such as giving a representation of polysemy through vectors (Jorge-Botana, León, Olmos, & Escudero, 2011), as a tool to enhance the comprehensibility of hypertext systems (Madrid & Cañas, 2011), or evaluating summaries from narrative and expository texts (León, Olmos, Escudero, Cañas, & Salmerón, 2006).

This paper focuses on Automatic Essay Evaluation and, especially, on automatic LSA assessments of student's summaries. We compare two LSA-based evaluation methods: Golden Summary and Inbuilt Rubric. LSA Golden Summary consists of comparing the vector representation of a text written by a study participant with the vector representation of one or more texts written by experts (e.g., Foltz, Laham, & Landauer, 1999; Klein, Kyrilov, & Tokman, 2011; Landauer, Foltz, & Laham, 1998). A single grade is obtained, as a function of the semantic distance between the student's text and the expert criterion. Golden Summary has frequently been used to evaluate summaries but a major limitation of this method is that the vector representation of a student summary does not capture the main ideas included in its answer; rather all the ideas are collected in a single vector (Olmos, Jorge-Botana, León, & Escudero, 2014; Olmos, Jorge-Botana, Luzón, Cordero, & León, 2016). To solve this limitation, Franzke, Kinstch, Caccamise, Johnson, and Dooley (2005) proposed an elaboration of partial golden summaries designed to detect students' misconceptions via thresholds, but the problem remains because the student's vector is conceived as a whole with the consequence that it is not easy to detect different ideas in the student's vectors. Moreover, creating partial golden summaries is time consuming and effortful. Another limitation of this method is that even when the perfect summary is redacted by an expert, the summary contains some level of bias towards one subject or another (Kintsch et al., 2000).

The Inbuilt Rubric method is a new method that accommodates a conceptual rubric in the LSA in order to detect contents more precisely and to overcome the limitations of the Golden Summary methods (Olmos et al., 2016). This method identifies the main contents of a text. In the first place, a rubric is elaborated by different experts where the main concepts of the instructional text are extracted (to ease the explanation, suppose it extracted *k* main concepts). After that, some lexical descriptors are provided to LSA to represent each of the *k* main conceptual ideas chosen previously. The *k* main concepts are represented in the LSA semantic space as *k* vectors. The last step consists of transforming the original latent semantic space to a semantic space where the first *k* dimensions have the meaning of the main concepts of the instructional text (a complete explanation of the method can be seen in Olmos et al., 2014; Olmos et al., 2016). Thus, the idea of the Inbuilt Rubric method is that the original semantic latent space, where the dimensions are meaningless is transformed into a new semantic space whose first *k* dimensions now capture the conceptual axes of the rubric. The main *k* ideas can be measured (quantified) in the student's vector summary as it is represented or projected in this new meaningful space. The student's vector summary is no longer represented as undifferentiated as in the Golden Summary method.

In addition to our interest in comparing the Golden Summary and Inbuilt Rubric approaches to applying LSA to the analysis of summaries, we were also interested in dimensions of rubrics that might affect how well the Inbuilt Rubric method performed. Because rubrics let users describe the characteristics of a product, project or text, they can be useful to teaching, learning, and assessment when they are well-designed (Dornisch & McLoughlin, 2006). Analytic rubrics list criteria to be assessed in student products (in this case, summaries) (Nitko, 2004) and let the evaluator provide feedback in order to improve the learning process of the student (Moskal, 2000). Some authors (e.g., Jonassen, Peck, & Wilson, 1999) have summarized the characteristics of a good rubric (for example, discrete criteria categories), while others (e.g., Tierney & Simon, 2004) have focused on factors that affect the rubric negatively (e.g., rubric descriptors that are either too general or too specific). One aspect of a rubric that may influence its effectiveness is the number of lexical descriptors per conceptual axis that are included in the rubric (Dornisch & McLoughlin, 2006). Thus, in the current study, we used few descriptors (three per axis) or many descriptors (5-8 per axis) to determine possible differences.

Other studies have found that some students write summaries with many irrelevant words (e.g., Olmos et al., 2016). For this reason, we introduced a second condition in the Inbuilt Rubric method: we compared a weighted and a non-weighted version of this method. As it was mentioned previously, *k* is the number of conceptual dimensions that is provided by the Inbuilt Rubric method. In the weighted version each of the *k* dimensions are multiplied by a W index. The W index is defined as:

$$W_i = inT_i / offT_i$$

where inT_i is the average score of the *k* conceptual dimensions of the student's *i* summary and $offT_i$ is the average scores in the remaining abstract (not conceptual) dimensions. While inT_i represents a measure of the relevance in the student's summary (information related with the conceptual axes), $offT_i$ is a measure of the information in a summary not related with the conceptual axes (irrelevant information to the topic). Thus, a high W value represents a summary that includes relevant, technical, and conceptual words (high inT_i), and at the same time avoids non-technical words or off-topic words (low $offT_i$). This W index prevents Inbuilt Rubric method from assigning a high score if a summary contains irrelevant ideas.

These four Inbuilt Rubric versions (few/many descriptors x weighted/unweighted method) were compared with the Golden Summary method for two different texts and two different student samples to analyze if there were differences in the reliabilities. Some studies found that LSA assessments were not precise enough when the number of words in a document was lower than 200 (Redher, Schreiner, Wolfe, Laham, & Kintsch, 1998), while other studies found that there were no differences with lower length summaries (León et al., 2006). In that way, the High School student sample was asked for a shorter summary (approx. 50) while the university sample was asked for a longer one (approx. 250 words).

With the goal to gain complementary evidences about the performance and the factors that affect LSA assessments, the sample was subdivided by the quality of the student summaries. As psychometric theory has established (e.g., Item Test Theory), measurement error depends on the level of the examinee's ability (e.g., Hambleton & Swaminathan, 2013). An interesting question was to analyze if there were differences in the LSA reliabilities methods in different ability groups, with the aim of studying for whom the proposed methods are most appropriate. Assuming different quantities of knowledge in each group, it was expected to find differences in the LSA performance in better and worst summaries. If consistent differences were found in the methods and experimental manipulations, LSA performance could be analyzed in order to

improve and to standardize the procedure establishing specific parameters.

The novelty of this study is to test the Inbuilt Rubric method in different experimental conditions to provide evidence about its assessments using a classical method (Golden summary method) as a baseline. This Inbuilt Rubric method let the user detect specific knowledge transforming the latent semantic space into a space with a semantic meaning. This study will try to inspect the LSA assessments' performance depending on the parameters that are used in the method (that is, the number of descriptors per axis and a weighting by the abstract dimensions).

In short, the aim of this study was twofold: a) to compare two automatic assessment methods using LSA for a student's assessment summaries — a classical method, Golden Summary (e.g., Foltz et al., 1999; Klein et al., 2011; Landauer et al., 1998) in which each student's summary was compared to an ideal summary created by experts, and Inbuilt Rubric (Olmos et al., 2014; Olmos et al., 2016), that accommodates an expert rubric through lexical descriptors into a LSA semantic space transforming these LSA space in such a way that the first dimensions captures the meaning of the rubric; b) to compare these two automatic assessment methods into two expository texts that differ in complexity (complex/easy), length (1,300/500 words), different readers (University/High School students), and quality of summaries following the criteria from human graders. To validate Inbuilt Rubric, four versions of the rubric (that is, the combination of few/many lexical descriptors and weighted/not weighted by abstract dimensions) were elaborated and were analyzed according to the quality of the evaluated summaries.

Method

Participants

A total of 166 subjects participated in this study. There were 81 undergraduate Spanish students from the Autonoma University of Madrid (59% female, average age 22.4) who read and summarized an expository text (*Darwin evolution theory,* 1,300 words length). Also, 85 participants were Spanish High School students (51% female, average age 14.4) who read and summarized a different, shorter expository text (*Strangled Trees,* 500 words length).

Materials

Texts. The instructional text presented to the university sample was about *Darwin's theory of evolution* (1,300 words); the text presented to the secondary school sample was about *Strangled Trees (500 words).* Darwin's text was an extract from Isaac Asimov's Great Ideas of Science (1969) while *Strangled trees* text was an extract from a Science textbook (Peiro, 1972). Both texts had an appropriate and coherent discourse about their contents, as well as appropriate to each student group level. Both were expository texts, which have shown better results in LSA assessments compared to other types of text such as narratives (León et al., 2006; Wolfe, 2005).

Corpus. A general domain corpus extracted from the Spanish Wikipedia composed of digitalized texts was used as the training corpus (404,436 documents and 39,566 unique terms) for both texts. The weighted function used was log-entropy (Nakov, Popova, & Mateev, 2001). A total of 300 dimensions were imposed for the latent semantic space.

Software. Both the training and the ensuing change of basis and re-orthogonalization of the space were carried out using Gallito 2.0 (Jorge-Botana, Olmos, & Barroso, 2013), software that makes it possible to perform the entire Inbuilt Rubric method process.

Procedure

Eight PhD psychology students attended a seminar (4 sessions, 8 hours) in which they learned to summarize and to evaluate the knowledge or semantic content in a text using a shared criteria stablished by their own. With the aim to redact an ideal summary that showed all the semantic content of the text, four of this PhD students independently created summaries of 250 words for the Darwin text; the other four PhD students independently created summaries of 50 words for the Trees text. These ideal summaries were created in order to extract the conceptual axes of the text and, also, as the input for the Golden Summary method. This first step was conducted to establish a good baseline in the Golden Summary method in order to have a reliable measure with which the Inbuilt Rubric method could be compared. In this way, the quality of the evaluation with the Golden Summary method was the principal concern and this baseline was carefully established in order to have good reliability with which we could compare human assessment of the summaries.

Once the PhD students' ideal summaries were created, the undergraduate university group was asked to read the Darwin text and to create summaries of 250 words from it, while the High School students group was asked to create summaries of 50 words from the Trees text.

Expert judges' rubrics of the two texts. Four different expert judges (the PhD students) assessed the summaries of each text on a 0 to 10 scale. This assessment was established by a rubric that contained the conceptual dimensions of the text (five conceptual axes for the Darwin text and four for the Trees text). The expert judges' rubrics of both texts were the result of the discussion of the judges about the ideal summaries that each of them created independently. From those discussions, the expert judges created a rubric that contained the common information that was present in every ideal summary. The evaluation of the student summaries was completed by the expert judges before any LSA assessment was carried out.

In the case of the Darwin text, four judges created a rubric to assess the quality of a student's communication of the text's main concepts, assigning 0 to 10 points to a summary. The Darwin text's conceptual axes were "earth's age" (maximum score = 2 points), "Lamarck" (max = 2), "Darwin's expedition" (max = 2), "Darwin's theory" (max = 3), and "transcendence of the theory" (max = 1). Each of the conceptual axes score given by the judges were summed to compose a final score (min = 0 and max = 10). The four judges' reliabilities for their scoring of the Darwin text ranged between .89 and .93 (Pearson correlation).

The analogous procedure was followed in creating a rubric for the Trees text. There were four conceptual axes for this text. The first conceptual axis referred to the proper localization of strangler trees (e.g., jungle areas, tropical areas, etc.). The second conceptual axis consisted of the description of the process of strangulation by means of the roots. The third conceptual axis was the fierce competition of the trees for reaching sunlight in the dense jungle. Finally, the fourth conceptual axis had to do with a general strategy of survival in difficult adaptation conditions. For the Trees text, another four judges independently assessed each of the 85 summaries written by the students. As in the Darwin text, each judge had to assign 0 to 2 points to each of the four conceptual axes created when establishing the rubric (not necessarily integer values) in order to compose a final score (min = 0 and max = 8). The reliability among the four judges ranged between .78 and .94 (Pearson correlation).

As the summaries created by the expert judges were the basis for extracting the conceptual axes of the text and, also, as the input for the Golden Summary method, measures in the Golden Summary and the Inbuilt Rubric methods were equivalent in the knowledge that they contained. The Golden summary method transforms the student summary into a vector and compares it with the expert judge summary vector, giving the similarity (similar to correlation) between both vectors as the assessment. The Golden Summary assessment

Table 1. Lexical Descriptors per Dimension (Conceptual Axis) in the *Darwin* Text

Conceptual axis	Few lexical descriptors			Many lexical descriptors				
Earth's age	Hutton	Buffon	earth	million	Years	planet	Lyell	
Lamarck	Lamarck	characteristics	acquired	giraffes	antelopes	effort	zoological philosophy	
Darwin's expedition	Beagle	Galapagos	finches	journey	ocean	Pacific	beaks	seeds
Darwin's theory	selection	natural	evolution	Malthus	modifications nature	survival	specialization	evolutionary advantages
Transcendence of the theory	polemic	biology	modern	revolution				

Note. For each conceptual axis, there were two versions of the lexical descriptors: a) few (3) descriptors and b) many descriptors (5-8) which were composed by the union of those few descriptors adding more lexical descriptors per axis. Original descriptors were written in Spanish.

Table 2. Lexical Descriptors per Dimension (Conceptual Axis) in the *Strangler Trees* Text

Conceptual axis	Few lexical descriptors			Many lexical descriptors			
Contextualization of the text	tree	strangle	Brasil	jungle	humid	tropical	
Process of strangulation	kill	asphyxia-tion	roots	epiphyte	choke	sap	host
Competition of the trees for reaching sunlight	competition	lights	sun	growth	forest	dark	
Strategy of survival in difficult adaptation conditions	adaptation	survival	survive	efficacy	biological	habitat	

Note. For each conceptual axis, there were two versions of the lexical descriptors: a) few (3) descriptors and b) many descriptors (6-7) which were composed by the union of those few descriptors adding more lexical descriptors per axis. Original descriptors were written in Spanish.

was established using the mean of the similarity between the student summary and each of the expert judge's summaries.

In the case of Inbuilt Rubric method, a new latent semantic space was generated, where the first dimension carried the meaning of each conceptual axis (five for the Darwin text and four for Trees text; see Olmos et al., 2014 or Olmos et al., 2016, for a complete description of this method). For the evaluation of the summaries, each summary was projected in this new semantic space and the coordinates from the dimensions with meaning were added in order to obtain a total grade. Two different variables, each one with two conditions, were analyzed in Inbuilt Rubric method. This resulted in a total of four combined conditions. The first manipulated variable was the number of lexical descriptors. As each of the conceptual axes was projected into LSA vector space, it was studied whether the number of lexical descriptors resulted in different reliabilities. Thus, a condition called *few descriptors* for each lexical descriptor was analyzed. In this case, a maximum of three descriptors was used to project the conceptual axe into LSA. The other condition was called *many descriptors*, where a maximum of eight was used to make the projection. Both few and many descriptors conditions for each text can be seen in Tables 1 and 2. The second manipulated variable was the weighting of the Inbuilt Rubric method for abstract dimensions or not weighted for abstract dimensions. As the Inbuilt Rubric method transforms the first p latent dimensions into meaningful dimensions (p is the number of conceptual axes), the remaining k–p dimensions are latent (abstract dimensions; note that in our case k, the total number of training dimensions, were 300). In the weighted version of Inbuilt Rubric, the final score is calculated as the addition of the meaningful LSA scores (in the p first dimensions) divided for the average of the absolute scores in the abstract dimensions (in the k–p dimensions). The idea of the weighted versions is to penalize those student summaries that have high score in the irrelevant (abstract or latent) dimensions because it is supposed that they lead to non-relevant information. The non-weighted version of Inbuilt Rubric simply calculates the score of a summary as the sum of the meaningful LSA scores (in the p first dimensions).

Data Analysis

Expert judges' assessments were calculated as the average of the total mark of every judge (which was calculated as the sum of the human evaluations in the text conceptual axis) in order to gain reliability avoiding bias towards single expert judges' assessments. The reliability of the assessments of both automatic methods (Golden Summary vs. Inbuilt Rubric) was calculated as the Pearson's correlation coefficient between the method's assessment and the mean of the expert judges. To analyze differences between the reliabilities of the automatic LSA methods, X^2 difference tests were conducted (via nested models). Also, to have a deeper understanding of how well the LSA assesses summaries, the sample was divided in three equal groups by level of performance using the original expert judges' assessments. Then, reliability was calculated in each of the three groups.

Results

Intergrades Agreements

The Trees text was summarized by High School students, who had a mean of 50 words per summary (from min 9 to max 124 words, *SD* = 24.4), while the Darwin text was summarized by students from a higher academic level with a mean of 185 words per summary (from min 48 to max 299 words, *SD* = 66.0). First of all, reliabilities were calculated among the human experts as it was the criteria to assess the LSA methods. The *intraclass correlation coefficient* between the four human experts in the *Strangler Trees* instructional text (N = 85 summaries) was .816. Thus, reliability in this text was high. Moreover, the *intraclass correlation coefficient* found for the four human experts in the Darwin text (N = 81summaries) was .859. Thus, the summary assessments were similar among the experts in both texts. The criteria to compare and judge the LSA methods were reliable.

Comparing Human and LSA Methods Reliabilities for Each Text

First of all, an overall analysis was conducted to examine if there were differences in the reliabilities (Pearson correlation matrices) between the two texts. To do this, the likelihood ratio test was used (Raykov & Marcoulides, 2008, p. 430) to test the null hypothesis of no text differences in the five human-LSA reliabilities (Golden Summary

Table 3. LSA Methods Reliabilities in the *Strangler Trees* Text (as the Pearson's Correlation Coefficient between the Human and the LSA Assessments)

	Golden Summary	IR few descriptors (n = 3)	IR many descriptors (n = 5-8)	IRW few descriptors (n = 3)	IRW many descriptors (n = 5-8)
Human assessment	.47	.79	.77	.63	.60
Golden Summary		.41	.36	.79	.63
IR few descriptors			.99	.65	.67
IR many descriptors				.61	.69
IRW few descriptors					.87

Note. IR = Inbuilt Rubric; IRW = Inbuilt Rubric Weighted; n = number of descriptors; N = 85. All Pearson's correlation coefficients were significant at $p < .01$ (bilateral).

Table 4. LSA Methods Reliabilities in the Darwin Text (as the Pearson's Correlation Coefficient between the Human and the LSA Assessments)

	Golden Summary	IR few descriptors (n = 3)	IR many descriptors (n = 5-8)	IRW few descriptors (n = 3)	IRW many descriptors (n = 5-8)
Human assessment	.67	.61	.70	.78	.77
Golden Summary		.75	.72	.78	.74
IR few descriptors			.89	.86	.80
IRmany descriptors				.92	.94
IRWfew descriptors					.96

Note. IR = Inbuilt Rubric; IRW = Inbuilt Rubric Weighted; n = number of descriptors; N = 81. All Pearson's correlation coefficients were significant at $p < .01$ (bilateral).

and the four Inbuilt Rubric reliabilities studied). The analysis compared a model with the restriction of five equal reliabilities between the two texts to a model without this restriction. The chi-square test showed a significant degree of fit, $\chi^2(5) = 21.235$, $p = .0007$. Thus, the null hypothesis of text equality in the reliabilities was rejected. Table 3 and Table 4 show the sample reliabilities between human and LSA methods in the Trees and Darwin texts, respectively. As will be shown later, the Golden Summary method does not perform as well as the Inbuilt Rubric method. Considering only Inbuilt Rubric methods, the main difference between the two texts was in the weighted reliabilities: the weighted reliabilities for the Darwin text were significantly higher than the weighted reliabilities in the Trees text ($p = .047$).

We also analyzed if the difference between the average of the four Inbuilt Rubric reliabilities and Golden Summary reliability was the same for the two texts. As can be seen in Table 3 and Table 4, the sample differences between Inbuilt Rubric reliabilities and Golden Summary reliability are higher for the Trees text. The null hypothesis of equal differences between the two texts was marginally significant ($p = .051$). Integrating these results with the results of the overall analysis, there is a significant interaction effect between the text and the human and LSA reliabilities and it is necessary to see each text separated.

Regarding the Trees text (N = 85 High School students), Table 3 shows the correlation coefficients (reliabilities) between the different methods. It can be observed that the unweighted model versions had significantly higher LSA-human reliabilities than the weighted versions, $\chi^2(2) = 9.496$, $p = .009$. Moreover, the model with few descriptors seems to work better than an accommodated rubric with many descriptors, but in this case not in a substantive way. A substantive result is the Golden Summary method where its reliability is lower with respect to the Inbuilt Rubric method. For example, there were significant differences between the reliabilities of Inbuilt Rubric *few descriptors* ($r = .785$) and Golden Summary ($r = .471$), $\chi^2(1) = 78.41$, $p < .001$, and also between Inbuilt Rubric *many descriptors/weighted* ($r = .601$) and Golden Summary ($r = .471$): $\chi^2(1) = 37.93$, $p < .001$.

Regarding the Darwin instructional text (N = 81 university students), Table 4 shows that weighted versions (with *few* and *many* descriptors) have higher correlations. We analyzed whether there were significant differences between the reliabilities of the Inbuilt Rubric and the Golden Summary methods. Differences were found between Inbuilt Rubric weighted with *few* descriptors and Golden Summary methods, $\chi^2(1) = 5.796$, $p = .016$,, and also between Inbuilt Rubric weighted with *many* descriptors and Golden Summary methods, $\chi^2(1) = 4.160$, $p = .041$. Reliabilities between the unweighted versions and Golden Summary did not reach significance.

Comparing LSA Methods Reliabilities about the Quality of Summaries

To determine the robustness of the LSA assessments and following the initial assessment of the expert judges as criteria, the summaries from the participants were divided in three groups for each text, resulting in summaries with low (33%), medium (33%), and high quality (33%).

LSA methods make a good assessment of the low-quality summaries (see Table 5). For the medium quality level, the Golden

Table 5. Reliabilities of Each Method for the *Darwin* Text

	Golden Summary	IR few descriptors (n = 3)	IR many descriptors (n = 5-8)	IRW Few descriptors (n = 3)	IRW many descriptors (n = 5-8)
Low quality	.53**	.59**	.60**	.69**	.66**
Medium quality	.28	.55**	.68**	.56**	.69**
High quality	.18	.28	.35	.57**	.48*

Note. IR = Inbuilt Rubric; IRW = Inbuilt Rubric Weighted; n = number of descriptors; N = 81.
*$p < .05$ (bilateral), **$p < .01$ (bilateral).

Table 6. Reliabilities (as the Pearson's Correlation Coefficient with Human Assessment) of each Method (Inbuilt Rubrics and Golden Summary) in the *Strangler Trees* Text

	Golden Summary	IR few descriptors (n = 3)	IR many descriptors (n = 5-8)	IRW few descriptors (n = 3)	IRW many descriptors (n = 5-8)
Low quality	.23	.64**	.62**	.33	.21
Medium quality	.01	.57**	.58**	.37	.56**
High quality	.21	.38	.36	.55**	.50**

Note. IR = Inbuilt Rubric, IRW: Inbuilt Rubric Weighted; n = number of descriptors; N = 85.
**p < . 01 (bilateral).

Summary method does not have a significant Pearson's correlation coefficient while the Inbuilt Rubric works fine. When the summaries with the higher quality were assessed, only the weighted versions of the Inbuilt Rubric obtained a significant Pearson's correlation coefficient with the human assessment.

For the Trees text there is less consistency than for the Darwin text (see Table 6). First, the Golden Summary method does not have a significant Pearson's correlation coefficient in any of the quality groups. Unweighted versions of the Inbuilt Rubric method worked better for the low and medium groups while the weighted versions have higher Pearson's correlation coefficient for the high-quality groups.

It is noteworthy that because we have divided the High School and university samples in three level groups (33% each), the restriction of range in the quality of summaries causes the obtained reliabilities to be lower. However, although all the methods were affected, there are actual and evident differences between the methods. As far as we see, the LSA Inbuilt Rubric seems to have good criteria in the discrimination of summary quality in all the levels (although we will discuss the differences between the Inbuilt Rubric versions), and thus the Inbuilt Rubric is more robust to the restriction of these range limitations.

Discussion

The Inbuilt Rubric method represents a new LSA strategy to extract the main concepts from a text. This new method transforms the original latent semantic space into a meaningful one (into a human expert rubric) where its first dimensions best represent the main concepts of a text. However, although the Inbuilt Rubric method has been previously applied (Olmos et al., 2016), it is not well understood yet which are the best parameters needed to represent an expert rubric into the meaningful transformed space. Thus, an empirical study was conducted with two samples where four versions of Inbuilt Rubric method were created by combining the number of descriptors to generate the rubric into the LSA space (few and many) and weighting (or not) the student vector summaries by the abstract dimensions. Also, the Inbuilt Rubric method was compared with another classic LSA assessment method (the Golden Summary method) taken as a baseline. In order to gain generalization of the comparison of both methods and the analysis of the parameters, two different expository texts and samples were used asking them to produce summaries with different number of words. Using the similarity of the LSA assessments with the human evaluation as the reliability criteria, results showed a higher reliability of the Inbuilt Rubric and, especially, of the weighted version of it with few descriptors.

First of all, we conducted an overall analysis to study the interaction effect between the text type (Darwin and Trees texts) and the human-LSA reliabilities. In general, the Inbuilt Rubric method obtained higher reliabilities than Golden Summary baseline method in both texts when the global evaluations were compared. This result points to a higher performance of the Inbuilt Rubric compared with the Golden Summary because the characteristics of the summaries used in this study were favorable for the Golden Summary method (besides the fact that it was used as a good baseline to compare with it the Inbuilt Rubric method assessments). That is, using more than one expert summary in order to avoid the drawback of having some level of bias towards one subject or another in the expert summary (Kintsch et al., 2000) and, also, using summaries with a length of more than 200 words in the Darwin text, which has been considered as the optimal length for this method (Redher et al., 1998). An explanation of the poor results of the Golden Summary method is that it produces a unique vector representation of the summary and does not assess the main ideas or concepts included in the summary. The Inbuilt Rubric method works in a different way: it captures each of the conceptual dimensions included in a summary because this method implements or projects a rubric into a new meaningful semantic space that has been previously established by the human experts (Olmos et al., 2014; Olmos et al., 2016).

The manipulation of the number of lexical descriptors per LSA axis (few/many) did not show statistically significant differences, although some authors (e.g., Tierney & Simon, 2004) noted that human rubric descriptors that are either too general or too specific are a factor that affects the quality of the rubric assessment. This result means that a higher specificity of the LSA dimensions does not necessarily produce higher reliabilities and that, in general, three lexical descriptors per axis seems to be a good option because more descriptors do not create better results. Since evaluators have devoted a great amount of time to creating rubrics that are effective educational instruments (Dornisch & McLoughlin, 2006), it could be an unnecessary effort for the user of this method to create a more complex version of the LSA rubric.

Concerning the weighted vs. not weighted version, a significant interaction effect showed the superiority of the weighted version in the Darwin text over the Trees text. The weighting procedure is intended to penalize those summaries that have a great amount of irrelevant information (redundant or non-technical information). As it was described, the weighted version consists of dividing the LSA's conceptual or meaningful axes of the rubric (p dimensions) by the abstract dimensions (the k - p dimensions) (Olmos et al., 2016). The hypothesis that could explain why the weighted version works especially well in the Darwin text is because the length of these summaries is considerably larger than the Strangled Trees summaries. In the Strangled Trees summaries, the weighted version could not detect well the irrelevant information because they are not long enough and the abstract dimensions cannot extract off-topic information.

Regarding to the quality of summaries using human assessment as criteria, there were differences between Golden Summary and Inbuilt Rubric methods, with the Inbuilt Rubric performing better. But there were also differences within Inbuilt Rubric versions depending on the texts. Although the division by quality creates some range constraint, the differences and comparisons observed between both methods (and the manipulations of the Inbuilt Rubric method) allowed a finer analysis of why they worked in that way in the whole sample. This detailed analysis showed that all methods are able to discriminate between low-quality summaries. However, in high-quality summaries the Inbuilt Rubric method (especially the weighted version) was the only one able to discriminate between them. This result shows that the Inbuilt Rubric method is able to discriminate the amount

of knowledge in all of the quality groups. In this way, some authors (Graesser et al., 2000) found that classic LSA assessments are more similar to intermediate experts than to more accomplished experts in the field, which means that the LSA can discriminate between better and worse summaries, but only a skilled method (like the new Inbuilt Rubric) can obtain finer assessments.

Since the results were analyzed in two different and heterogeneous samples, differences between Golden Summary and Inbuilt Rubric methods seem to be very consistent. The main limitation of the LSA studies is the use of a unique general corpus, which determines the results. The Wikipedia corpus used in this study to provide knowledge to LSA has a limited amount of knowledge about some concepts (for example, strangler trees) and, at the same time, it is challenging to correctly discriminate between close concepts (like Darwin and his theory of evolution) using such a general corpus. It is possible, however, that by using a more specific corpus (for example, a biological or an evolutionary one) the reliability and validity obtained would be higher as these corpuses may better distinguish between interrelated concepts than a general content domain corpus. Future research should try to find differences in the performance of the LSA assessments depending on the characteristics of the linguistic corpus that is used to train the LSA.

Although the characteristics of the weighted Inbuilt Rubric method were selected in order to analyze the relevant vs. irrelevant information included in a student summary, in Olmos et al. (2016) the weighted version also took into account the number of words of a summary (e.g., penalizing summaries with excessive number of words). Thus, other weighted versions should include features as, for example, number of words or redundancy and not only in-topic and off-topic characteristics of a summary.

A recent study involving a sample of 864 university students demonstrated high ecological validity as 85% of the students expressed satisfaction with the feedback provided by the Inbuilt Rubric method (Olmos et al., 2016). These students perceived the method as useful for improving their text comprehension. The development of automatic assessment methods like Inbuilt Rubric holds great promise as tools that will guide students to improve their performance in reading and writing skills as well as their capacity of summarizing (Foltz, Gilliam, & Kendall, 2000).

As it was presented in this study, the Inbuilt Rubric method simulates human assessment better than the Golden Summary independently of the complexity or length of the text and the academic level of the reader. The Inbuilt Rubric method transformed a latent space into a topic or meaningful semantic space. As the dimensions represent concepts from a rubric, Inbuilt Rubric detected which contents are or not included in a student summary without creating partial Golden Summaries, which was an alternative created to detect specific topics in a text but its costs in terms of time and effort were very high (Olmos et al., 2016). Future research should analyze the Inbuilt Rubric method in order to continue the standardization of its parameters and should develop new applications that would let users improve their skills without assuming high costs. Another goal would be generalizing these results to new texts that have different characteristics (for example, narrative texts) in order to improve current LSA applications.

Conflict of Interest

The authors of this article declare no conflict of interest.

Acknowledgments

This paper is especially dedicated to our friend and workmate, Adrian Mencu. The authors would like to thank Robert F. Lorch for comments and revision of an earlier version of this manuscript.

References

Aryal, A., Gallivan, M., & Tao, Y. (2015, August). *Using Latent Semantic Analysis to Identify Themes in IS Healthcare Research.* Paper presented at the Twenty-first Americas Conference on Information Systems (AMCIS).

Asimov, I. (1969). *Great Ideas of Science.* Boston, MA: Houghton Mifflin.

Borisov A., Serdyukov P., & de Rijke, M. (2016, April). *Using metafeatures to increase the effectiveness of latent semantic models in web search.* Paper presented at the 25th World Wide Web Conference (WWW 2016) (pp. 1081-1091). https://doi.org/10.1145/2872427.2882987

Cohen, T., Blatter, B., & Patel, V. (2008). Simulating Expert Clinical Comprehension: Adapting Latent Semantic Analysis to Accurately Extract Clinical Concepts from Psychiatric Narrative. *Journal of Biomedical Informatics, 41,* 1070-1087. https://doi.org/10.1016/j.jbi.2008.03.008

Dornisch, M. M., & McLoughlin, A. S. (2006). Limitations of web-based rubric resources: Addressing the challenges. *Practical Assessment, Research & Evaluation, 11*(3), 1-8.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interacting Learning Environments, 8*(2), 111-129. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT111

Foltz, P. W., Laham, D., & Landauer, T. (1999). *Automated Essay Scoring: Applications to Educational Technology.* Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA '99) (pp. 939-944). Seattle.

Franzke, M., Kinstch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street: computer support for comprehension and writing. *Journal of Educational Computing Research, 33*, 53-80. https://doi.org/10.2190/DH8F-QJWM-J457-FQVB

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive learning environments, 8,* 129-147. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT129

Günther, F., Dudschig, C., & Kaup, B. (2015). Latent Semantic Analysis cosines as a cognitive similarity measure: Evidences from priming studies. *Quarterly Journal of Experimental Psychology, 69*, 626-653. https://doi.org/10.1080/17470218.2015.1038280

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications.* Berlin/Heidelberg, Germany: Springer Science & Business Media.

Jonassen, D. H., Peck, K. C., & Wilson, B. G. (1999). *Learning with technology in the classroom: A constructivist perspective.* New York, NY: Merrill/Prentice-Hall.

Jorge-Botana, G., León, J. A., Olmos, R., & Escudero, I. (2011). The representation of polysemy through vectors: some building blocks for constructing models and applications with LSA. *International Journal of Continuing Engineering Education and Long Learning, 21*, 328-342. https://doi.org/10.1504/IJCEELL.2011.042791

Jorge-Botana, G., Olmos, R., & Barroso, A. (2013, July). *Gallito 2.0: a Natural Language Processing tool to support Research on Discourse.* Proceedings of the Twenty-third Annual Meeting of the Society for Text and Discourse. Valencia.

Jorge-Botana, G., Olmos, R., León J. A. (2009). Using LSA and the predication algorithm to improve extraction of meanings from a diagnostic corpus. *Spanish Journal of Psychology, 12*, 424-440. https://doi.org/10.1017/S1138741600001815

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8*, 87-109. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT087

Klein, R., Kyrilov, A., & Tokman, M. (2011). *Automated assessment of short free-text responses in computer science using latent semantic analysis.* Presented in Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education (ITiCSE '11) (pp. 158-162). Darmstadt, Germany. https://doi.org/10.1145/1999747.1999793

Landauer, T. K. & Dumais, S. (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/10.1.1.184.4759

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

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *The Handbook of Latent Semantic Analysis.* Mahwah, NJ: Routledge.

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

Madrid, R. I., & Cañas J. J. (2011). Using latent semantic analysis to enhance the comprehensibility of hypertext systems. *International Journal of Continuing Engineering Education and Life-Long Learning, 21,* 343-354. https://doi.org/10.1504/IJCEELL.2011.042792

Moskal, B. M. (2000). Scoring Rubrics: What, When and How? *Practical Assessment, Research & Evaluation, 7*(10), 71-81.

Nakov, P., Popova, A., & Mateev, P. (2001, September). *Weight functions impact on LSA performance.* Presented at the EuroConference Recent Advances in Natural Language Processing (RANLP'01) (pp. 187-193). Sophia, Bulgary.

Nitko, A. J., (2004). *The educational assessment of students* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Olmos, R., Jorge-Botana, G., León, J. A., & Escudero, I. (2014). Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis. *Discourse Processes, 51*, 494-510. https://doi.org/10.1080/0163853X.2014.913416

Olmos, R., Jorge-Botana, G., Luzón, J. M., Cordero, J., & León, J. A. (2016). Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system. *Information Processing & Management, 52,* 359-373. https://doi.org/10.1016/j.ipm2015.12.002

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Pakhomov, S. V., & Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex, 55*, 97-106. https://doi.org/10.1016/j.cortex.2013.05.009

Peiro, A. (1972). *Ciencias de la Naturaleza 6º EGB.* Madrid, España: Anaya.

Pu, X., Jin, R., Wu, G., Han, D., & Xue, G. R. (2015). *Topic modeling in semantic space with keywords.* Presented at the 24th ACM International Conference on Information and Knowledge Management (pp. 1141-1150). Melbourne, Australia. https://doi.org/10.1145/2806416.2806584

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis.* New York, NY: Routledge.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25,* 337-354. https://doi.org/10.1080/01638539809545031

Ryan, J. O., Kaltman, E., Mateas, M., & Wardrip-Fruin, N. (2015). *What we talk about when we talk about games: Bottom-up game studies using natural language processing.* Paper presented at the 10th International Conference on the Foundations of Digital Games (FDG). California, USA.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2), 1-10.

Visinescu, L., & Evangelopoulos, N. (2014). Orthogonal Rotations in Latent Semantic Analysis: An Empirical Study. *Decision Support Systems, 62,* 131-143. https://doi.org/10.1016/j.dss.2014.03.010

Wendy, S. T. W., How, B. C., & Atoum, I. (2014). *Using Latent Semantic Analysis to Identify Quality in Use (QU) Indicators from User Reviews.* Paper presented at The International Conference on Artificial Intelligence and Pattern Recognition (AIPR2014) (pp. 143-151). Kuala Lumpur, Malaysia: SDIWC Publications.

Wolfe, M. B. W. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 359-364.

Xu, H., Zeng, W., Gui, J., Qu, P., Zhu, X., & Wang, L. (2015). *Exploring similarity between academic paper and patent based on Latent Semantic Analysis and Vector Space Model.* Paper presented at 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (pp. 801-805). Zhangjiajie, China. https://doi.org/10.1109/FSKD.2015.7382045