Enhancing topic-detection in computerized assessments of constructed responses with distributional models of language

José Á. Martínez-Huertas[*1,2], Ricardo Olmos[3] & José A. León[4]

1 =
Email: josea.martinez@uam.es
(+34) 91 497 87 50
Universidad Autónoma de Madrid
Department of Psychology
Calle Iván Pavlov, 6
28049, Madrid, Spain

2 =
Universidad Pontifica de Comillas
UNINPSI Clinical Psychology Center
Calle Mateo Inurria, 37,
28036 Madrid, Spain

* Corresponding Author

3 =
Email: ricardo.olmos@uam.es
(+34) 91 497 85 86
Universidad Autónoma de Madrid
Department of Psychology
Calle Iván Pavlov, 6
28049, Madrid, Spain

4 =
Email: joseantonio.leon@uam.es
(+34) 91 497 85 86
Universidad Autónoma de Madrid
Department of Psychology
Calle Iván Pavlov, 6
28049, Madrid, Spain

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ABSTRACT

Usually, computerized assessments of constructed responses use a predictive-centered approach instead of a validity-centered one. Here, we compared the convergent and discriminant validity of two computerized assessment methods designed to detect semantic topics in constructed responses: Inbuilt Rubric (IR) and Partial Contents Similarity (PCS). While both methods are distributional models of language and use the same Latent Semantic Analysis (LSA) prior knowledge, they produce different semantic representations. PCS evaluates constructed responses using non-meaningful semantic dimensions (this method is the standard LSA assessment of constructed responses), but IR endows original LSA semantic space coordinates with meaning. In the present study, 255 undergraduate and high school students were allocated one of three texts and were tasked to make a summary. A topic-detection task was conducted comparing IR and PCS methods. Evidence from convergent and discriminant validity was found in favor of the IR method for topic-detection in computerized constructed response assessments. In this line, the multicollinearity of PCS method was larger than the one of IR method, which means that the former is less capable of discriminating between related concepts or meanings. Moreover, the semantic representations of both methods were qualitatively different, that is, they evaluated different concepts or meanings. The implications of these automated assessment methods are also discussed. First, the meaningful coordinates of the Inbuilt Rubric method can accommodate expert rubrics for computerized assessments of constructed responses improving computer-assisted language learning. Second, they can provide high-quality computerized feedback accurately detecting topics in other educational constructed response assessments. Thus, it is concluded that: (1) IR method can represent different concepts and contents of a text, simultaneously mapping a considerable variability of contents in constructed responses; (2) IR method semantic representations have a qualitatively different meaning than the LSA ones and present a desirable multicollinearity that promotes the discriminant validity of the scores of distributional models of language; and (3) IR method can extend the performance and the applications of current LSA semantic representations by endowing the dimensions of the semantic space with semantic meanings.

Keywords: Inbuilt Rubric; constructed responses; summaries; topic detection; Latent Semantic Analysis; Automated Summary Evaluation

1. INTRODUCTION

Latent Semantic Analysis (LSA) is a well-known computational linguistic model of meaning representation. After being exposed to hundreds of thousands of documents, LSA represents meaning in a reduced *k*-dimensional space. The semantic representations of LSA have a substantial theoretical background and have been used to develop useful applications for psychological and educational measurement (Kaur & Sasi Kumar, 2019; Landauer & Dumais, 1997; Landauer et al., 2007; LaVoie et al., 2020; McNamara, 2011; Saha & Rao, 2019), as well as for Natural Language Processing (NLP; e.g., Hewitt & Manning, 2019; Suleman & Korkontzelos, 2021). Traditionally, the cosine-based similarity has been used to analyze the quality of essays (i.e., of constructed responses) and it has been considered an adequate semantic representation of texts. But the problem with the cosine-based similarity is that text representation would depend on the representation of other texts (e.g., a "golden" or ideal response). This limits the use of all capabilities that LSA encompasses. Against this computational perspective, the Inbuilt Rubric method (IR; Olmos et al., 2014, 2016) is proposed here as an appropriate semantic representation model for topic-detection in computerized assessments of constructed responses, due to its capacity to endow semantic space coordinates with meaning. IR enables using LSA in a richer manner because, compared to the cosine-based method, it focuses on the interpretability of the dimensions (e.g., vector representation of constructed responses can be used in absolute terms, and not in relative terms).

Predictive-centered approaches are usually based on the analysis of validity in terms of maximum performance (e.g., using R^2, bias or error variance). This approach sorts the compared methods according to these maximum performance criteria. As against this, validity-centered approaches are usually based on testing theoretical predictions and specific hypotheses. The motivation of the present study is to compare the two computerized assessment methods for constructed responses in a topic-detection task, using a validity-

centered approach: the Partial Contents Similarity (PCS) and the IR methods. Specifically, computerized assessment of student summaries was used to evaluate the topic-detection quality of semantic representation of IR and PCS methods. Thereafter, we shall test the convergent and discriminant validity of IR and PCS method scores, when predicting concepts from a validity-centered perspective using human rates as a golden criterion.

1.1. Some limitations about topic-detection using the cosine-based similarity in distributional models

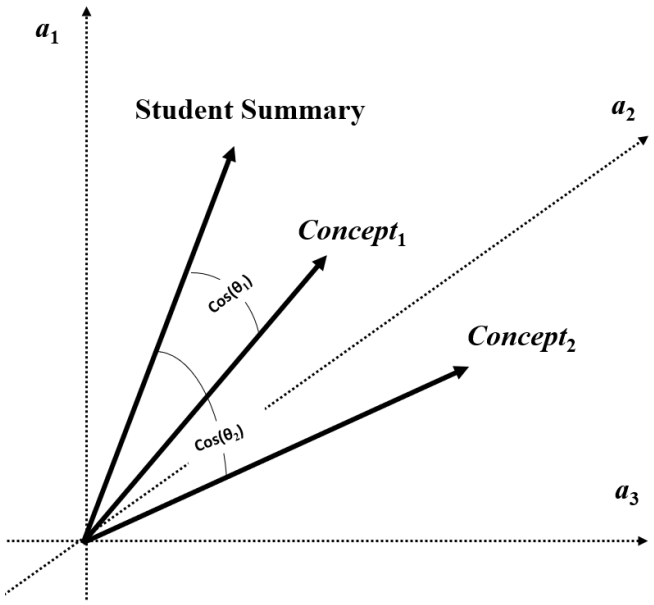
In distributional models of language, it is assumed that words that occur in similar contexts tend to have similar meanings (Deerwester et al., 1990). Then, LSA operationally proposed a cognitive mechanism that learns semantics from repeated episodic experiences in a linguistic environment (see semantic memory distributional models[1]). It generates a multi-vector semantic space transforming an initial matrix conformed to by the co-occurrence of words using *Singular Value Decomposition* (SVD) dimension reduction (Deerwester et al., 1990). It was originally proposed to extract and represent the semantic meaning of words to measure similarities among words and groups of words (Landauer & Dumais, 1997; Landauer et al., 2007). These classical LSA scoring methods compare vector representations of constructed responses with vector representations of gold-standard criteria, as a measure of text quality or similarity (León et al., 2006; Foltz et al., 1999; Klein et al., 2011).

A widely applied classic method for evaluating text concept representations is the cosine-based similarity of partial contents or PCS[2] (Dessus & Lemaire, 1999; Franzke et al., 2005; Kintsch et al., 2007; Magliano & Graesser, 2012). Here, the similarity of a text vector

[1] Distributional models of semantic memory assume the existence of a formal cognitive mechanism that learns semantics from repeated episodic experience in a linguistic environment. Distributional models are also known as corpus-based, semantic-space, or co-occurrence models (Jones et al., 2015). See also the recent revision by Günther et al. (2019).
[2] Partial Contents Similarity (PCS) method was originally conceived as a tool to detect specific contents in summarization tasks. Thus, it is also known as partial golden summaries.

representation (in this case, a student summary) is compared to another vector representation of a text sample (in this case, a fragment of the instructional text). This is the standard procedure used to evaluate the quality of constructed responses using LSA semantic representations. While this approach is very useful, a similarity measure (cosine) cannot properly represent texts, because vector representations in semantic space assemble different aspects of concepts, reducing vector representations to a simple comparison of strongly related inputs (Turney, 2006; Turney & Pantel, 2010; also known as the referential circle problem, de Vega et al., 2012). Other concerns refer to non-semantics, such as syntactic characteristics that influence cosine-based similarity (Kintsch et al., 2000), or the time-consuming and considerable efforts that require the generation of vector representation of the text samples (Dronen et al., 2015). *Figure 1* represents the evaluation of two concepts (C_1 and C_2) in a student summary. In this example, a unique graphical representation is made for both concepts, but the cosine-based similarity of the student summary with each concept must be computed separately. In this example, the student summary would have a richer semantic representation for the first concept (C_1) than for the second concept (C_2) as its cosine is higher.

Figure 1. Graphical representation of Partial Contents Similarity assessments for constructed responses (here, one student summary).



Note. Only three dimensions of the latent semantic space are represented to ease the interpretation (a_1-a_3). Here, two concepts ($Concept_1$ and $Concept_2$) are represented. The student summary that is represented have a higher cosine-based similarity with the first concept than with the second concept. Whilst the representation of the Partial Contents Similarity assessments is made jointly, these cosines must be computed separately.

It is worth mentioning here that automatic text analysis is usually discussed under the term *topic modeling,* which is based on the Latent Dirichlet Allocation (LDA) method (Blei et al., 2003). Both LDA and SVD methods (like LSA) process textual corpora differently, but they generate a specific semantic space whose concepts are naturally embedded within the information of the corpus. These methods extract some "topics" present in the textual corpora (it is to be noted that LSA does not extract explicit concepts: it just extracts abstract non-meaningful dimensions, as opposed to the LDA method, which is able to obtain explicit and meaningful topics). Thus, their extraction of concepts or topics is *a posteriori*, that is, they extract semantic concepts or topics defined by the contents of the textual corpora. Against this, the IR method transforms the LSA's latent semantic space using *a priori* imposed concepts or topics, that is, it extracts/represents some concepts or topics that users want to evaluate using a confirmatory perspective (see a more detailed explanation below).

1.2. A new "non-Latent" Semantic Analysis approach: Inbuilt Rubric (IR) method

Since classic similarity measures have considerable validity concerns relating to their capacity to represent concepts, a non-latent semantic space such as the one generated by the IR method becomes a useful space to analyze the potential of computational semantics. Here, the IR method is hypothesized to activate concepts in its semantic space to achieve specific task demands (in this case, assessing concepts in constructed responses). Moreover, although the performance of some computational models has been questioned when unrelated concepts are processed (De Deyne et al., 2016), the IR method makes an orthogonalization of task-related concepts to capture meaning and does not present any *a priori* dependency on the similarity of concepts[3].

The IR method was designed as a computational implementation of rubrics in LSA's space, using analytical and topic-specific scoring instead of holistic, since these properties are recommended for use in student-constructed response assessments (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). In this way, the overall scores of the IR method have attained higher performance compared to the classical LSA methods in different tasks, and its advantage has been attributed to its capacity to evaluate specific semantic contents (Martínez-Huertas et al., 2018, 2019). But it is necessary to analyze the validity of such meaningful semantic space, that is, its ability to detect what topics are present and lacking in student summaries.

Readers interested in the implementation of the IR method are referred to Hu et al. (2007) and Olmos et al. (2014). A brief summarization of its implementation follows. Once a dimension reduction procedure (such as SVD) is applied to a corpus as in other LSA methods,

[3] This is because IR uses an algebraic orthogonalization process (Gram-Schmidt) that makes the concepts independent; but, of course, the represented concepts need to be different enough to be orthogonalized before using the IR method.

the IR method transforms the latent semantic space into a non-latent semantic space. Thus, the two main steps of IR method implementation (described in Olmos et al., 2014) are:

1. Creating an algebraic basis $\boldsymbol{\beta}$, in which the main topics of a text must be incorporated using lexical descriptors[4].

   First, a classical LSA's latent semantic space, ***US***, is generated. Its dimensions are *n* unique words x *k* latent semantic dimensions. A standard dimensionality for LSA's ***US*** matrix is to impose *k* to be approximately around 300 dimensions (they are the *k* latent dimensions and the reason for choosing a number around 300 is fundamentally empirical; see, for example, Rehder et al., 1998). Then, it is necessary to identify the lexical descriptors that will form the new algebraic basis $\boldsymbol{\beta}$, where the first *p* vectors of $\boldsymbol{\beta}$ are the target concepts (see *2.4 section* for a more detailed explanation of lexical descriptor selection in the present study).

   $\boldsymbol{\beta}$ must have *k* x *k* dimensions (the first *p* forming meaningful concepts and the remaining *k* – *p* created by independent vectors, e.g., the canonical basis). Then, the algebraic Gram-Schmidt orthogonalization process is applied to the $\boldsymbol{\beta}$ matrix to orthogonalize and normalize it (it results in an orthonormalized basis that can be expressed by $\boldsymbol{\beta}^{ON}$). Consequently, its *k* vectors maintain linear independence[5]. For example, and to facilitate the understanding of this part, the $\boldsymbol{\beta}^{ON}$ matrix of one of the texts used in the present study was defined as $\boldsymbol{\beta}^{ON}=\{\mathbf{b}_{\text{Debate}}, \mathbf{b}_{\text{Phonology}}, \mathbf{b}_{\text{Syntax}}, \mathbf{b}_{\text{Semantics}}, \mathbf{b}_{\text{Symbol}}, \mathbf{b}_{\text{Abstract 1}}, \ldots, \mathbf{b}_{\text{Abstract k-p}}\}$. In this case, the first five dimensions are the target concepts and the rest *k* – *p* dimensions (i.e., 295) are independent vectors from the original basis.

[4] A lexical descriptor is a word that represents a concept. Specifically, several words are chosen (typically three or four) that best reflect the intended concept. For example, "fruit" is an accurate lexical descriptor of "strawberry".

[5] If two vectors between the first *p* meaningful ones are excessively correlated, then both concepts are not allowed to be part of the final basis $\boldsymbol{\beta}^{ON}$) and the concepts are rethought.

2. Transforming the latent semantic space into a meaningful one on this basis.

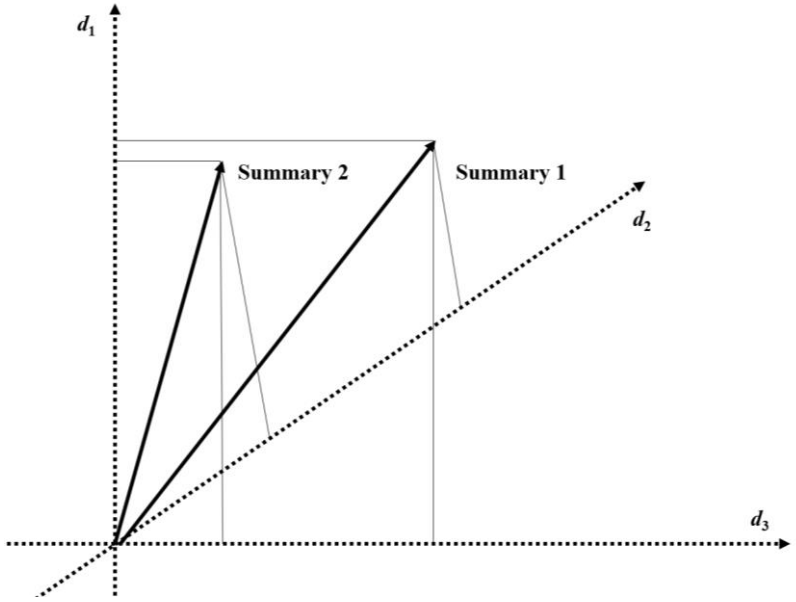
   The original LSA's ***US*** matrix is then expressed in the orthonormalized $\boldsymbol{\beta}^{ON}$ basis, generating a new term matrix ***C***, using the following formula:

$$\boldsymbol{C} = (\boldsymbol{US}) \cdot (\boldsymbol{\beta}^{ON})^{-1} \quad [1]$$

   where ***C*** is the meaningful semantic space with *n* unique words x *k* semantic dimensions. The key idea of all this process is that the first *p* dimensions of ***C*** are meaningful, that is, its *p* first dimensions can be interpreted as the semantic representations of these concepts.

Once this meaningful semantic space (***C***) has been generated, constructed responses are then projected onto it to evaluate them. *Figure 2* represents the evaluation of two student summaries. In this case, only three concepts are represented to ease the interpretation of this hypothetical example, but the IR method can evaluate different numbers of concepts. Summary 1 would have a higher general quality than Summary 2. The quality of the summary covering each concept can be seen as the projection on each of the dimensions (d_1, d_2, d_3) of this meaningful ***C*** semantic space. As can be observed in the graph, Summary 1 would have a higher score than Summary 2 in the second and third dimensions (d_2 and d_3, respectively), but they would have a similar score in the first dimension (d_1).

Figure 2. Graphical representation of Inbuilt Rubric assessments for constructed responses (here, student summaries).



Note. Only three meaningful concepts are represented to ease the interpretation (*d*1-*d*3). Summary 1 quality is higher than that of Summary 2 because projections onto meaningful semantic dimensions are larger. Summary 1 would have a higher score than Summary 2 in the second and third dimensions, but both would have a similar score in the first dimension.

In this way, the IR method would produce *k* scores: *p* scores (information related to the concepts) and other *k* – *p* LSA dimension scores (representing irrelevant information about the assessed topic). Then, the number of IR method scores would depend on the dimensionality of the original latent semantic space. Thus, since many student summaries usually include many irrelevant words (Olmos et al., 2016), IR method scores are weighted by a W_i index (where *i* represents each student). The W_i index is calculated using the following formula:

$$W_i = {}_{in}T_i \,/\, {}_{off}T_i \qquad [2]$$

where ${}_{in}T_i$ is the average of the *p* scores, and ${}_{off}T_i$ is the average score of the other LSA dimensions. A higher W_i index means that relevant or conceptual words are included, while a lower W_i index signifies the inclusion of many non-technical or off-topic words. This W_i index could discriminate between the quantity of relevant and irrelevant ideas contained in the evaluated constructed response.

1.3. Some implications of topic-detection in computerized assessments of constructed responses

As LSA semantic representations have been used for topic-detection in computerized assessments of constructed responses and thus for educational technology research (e.g., Landauer et al., 2007, including the PCS method), enhancing the quality of the semantic representations could improve its applications. Both methods are designed for topic-detection in the assessment of constructed responses and thus their usefulness is similar. For example, we think that the potential improvement of IR method could be relevant for the development of intelligent tutoring systems and other expert systems.

In the first place, the IR method could have straightforward implications for intelligent tutoring systems. For example, its semantic representations could improve current computer-aided interventions whose performance is comparable to human interventions (VanLehn, 2011) by providing high-quality feedback in intelligent tutoring systems, which can be a key aspect for enhancing learning (e.g., Kaur & Kumar, 2019; Roll et al., 2011). Moreover, these semantic representations can be useful for different summarization tasks from relevant institutions such as the *Educational Testing Service* (Madnani et al., 2013) or complementing other intelligent tutoring systems such as ElectronixTutor (Graesser et al., 2018). In the second place, the multi-vector semantic representations of the IR method could enhance some procedural phases such as data analysis and labeling, feature selection and projection, or solution evaluation (see the elements of text classification in Mirończuk & Protasiewicz, 2018; or the proposals by Jorge-Botana et al., 2019).

It is noteworthy that some computerized assessment methods that use neural networks or other machine-learning algorithms can obtain desirable results (e.g., He & Lin, 2016; Shen et al., 2014; Wang et al., 2015); but IR and PCS methods try to produce interpretable scores based on the meaning of their respective vector representations. In this way, the meaningful

scores of the IR method could improve many of the applications of LSA and other distributional models, especially when multi-vector representations are needed (Kundu et al., 2015). In fact, recent and interesting LSA applications, such as the one by Kundu et al., 2015, developed new procedures to interpret the meaning of LSA space applying varimax rotations. Thus, IR scores are a good alternative to extract semantic vector representations through interpretable multi-vector representations of constructed responses, following a more validity-centered than a predictive-centered approach.

1.4. The present study

In addition to the hypothetical cognitive mechanism that learns semantics from repeated episodic experiences in a linguistic environment (characteristic of distributional models), the IR and the PCS methods model the activation of some concepts in the semantic network. But both methods use qualitatively different vector representations in their respective semantic spaces: while the PCS method represents meaning as a similarity measure between one text vector representation and another, the IR method represents meaning in multi-vector representations based on orthogonal dimensions. As a working example, the same meaning of a summary would be represented as the cosine (a scalar score for each concept) in the PCS method and as a group of orthogonal scores in the IR method. In the present study, we analyzed the convergent and discriminant validity (Campbell & Fiske, 1959) of the semantic representations of PCS and IR computational methods for topic-detection in computerized constructed response assessments. As was stated by Campbell and Fiske (1959), this validational process consists in the evaluation of inter-correlations of different constructs measured by different methods. Ideally, the measures of the same construct should correlate higher than other constructs within the same method, compared to different constructs. Then, higher convergent and discriminant validity evidence require a method to correlate higher than other methods, when measuring a construct in comparison to non-measured constructs. Here, we analyzed the convergent and discriminant validity

between human raters and computational methods using human rates as a golden criterion, but also the convergent and discriminant validity between computational methods to analyze their potential similarities. In the first case, we expect to find a higher convergent and discriminant relation for the IR method, scores in comparison with PCS method scores. In the second case, we expect to find convergent and discriminant relations, as both methods use the same LSA's previous knowledge.

Previous research found that overall scores of the IR method outperform the classic methods in Automated Summary Evaluation (Martínez-Huertas et al., 2018, 2019; Olmos et al., 2016), but this is the first time that the coordinates have been tested to investigate their capacity to represent specific semantic contents. Thus, the aim of the present study was to analyze whether the IR method coordinates can represent concepts, improving the performance of widely applied procedures such as the PCS method. Then, both methods were tested to solve specific topic-detection task demands in constructed response assessments within Automated Summary Evaluation. Thus, whilst the IR and the PCS methods are distributional models of language and use the same LSA prior knowledge, they were expected to produce different semantic representations because the former uses a topic-based strategy to impose a priori semantic meanings in the semantic space. Apart of comparing the performances of IR and PCS methods based on convergent and discriminant validity, their computational scores were also analyzed in terms of their multicollinearity and their semantic meanings. It was expected to find evidence in favor of the meaningful multi-vector representations of IR method in front of the classic LSA measures as the ones of the PCS method. As it is later discussed in the light of the results of the present study, the meaningful semantic vector representations of IR method could enhance current computer-assisted language learning applications.

2. MATERIAL AND METHODS[6]

2.1. Participants

A total of 255 undergraduate (average age 21) and high school students (average age 17-18) from different institutions in Madrid summarized one text from a total of three texts. They were recruited as volunteers and received course credit for their participation in this study. Eighty-eight students summarized Text 1 (*Darwin's Theory of Evolution*), 76 students summarized Text 2 (*Strangler Trees*), and 91 students summarized Text 3 (*Theory of the Evolution of Language*).

2.2. Instruments and materials

2.2.1. Texts

Three expository texts were selected to provide higher experimental control because LSA methods tend to obtain a higher performance in expository texts (e.g., Wolfe, 2005; León et al., 2006). Texts were written in Spanish. Thus, we equated the levels of the experimental texts according to criteria established in the *Common European Framework of Reference for Languages* (CEFR) following the specific Spanish descriptors of each mastery skill from the *Curriculum Plan of the Cervantes Institute*.

Text 1 is *Darwin's Theory of Evolution* (Asimov, 1969), with a length of approximately 1,300 words and describes how Darwin was influenced by other authors and how he developed his theory of evolution. 4.8% of this long text consisted of technical words and its difficulty corresponded to level B2 in the CEFR, i.e., medium difficulty.

Text 2 is *Strangler Trees* (Peiro, 1972), with a length of approximately 500 words and describes how species of trees compete for alimentary recourses to survive. 2.4% of this short

[6] The instructional texts, the constructed responses (student summaries), and the data sets of the study can be found in the following OSF project: https://osf.io/yra7n/. The constructed responses and the human ratings can be used to test the performance of other computational methods used to evaluate constructed responses.

text consisted of technical words and its difficulty corresponded to level B1-B2 in the CEFR, i.e., low difficulty.

Text 3 is *Theory of the Evolution of Language* (Martín-Loeches, 2016), with a length of approximately 900 words and describes different theories of the evolution of language. 19.81% of this text consisted of technical words and its difficulty corresponded to level C1 in the CEFR i.e., high difficulty.

2.2.2. Assessment rubrics

The assessment rubrics for human raters were created using an inductive process. First, two human raters (Ph.D. students trained to summarize instructional texts) read the texts and summarized them with the aim of generating an ideal summary. Second, we systematically evaluated the essential information of those summaries to extract the main concepts of the text (common and necessary topics for good constructed responses were extracted by consensus in discussion groups). Then, we defined the assessment criteria of concepts following Jonsson & Svingby (2007) and León et al. (2006) procedures. Specifically, the assessment of concepts consisted of the consideration of the inclusion of some sub-topics and a coherent discourse. In general, rubric scores were scaled from 0 (omission of the concept) to 2 (coherent and full explanation of the concept) using 1 for partial presence of concepts (intermediate scores like 0.5 or 1.5 were also allowed). As can be observed in the *3.1. section*, this was an efficient and reliable way of measuring constructed responses. In this manner, contents that should be included in good summaries were used to compound three assessment rubrics.

The rubric for Text 1 was composed of five concepts: *Earth's age* (maximum score in the rubric = 2 points), *Lamarck* (max = 2), *Darwin's expedition* (max = 2), *Darwin's theory* (max = 3), and *Transcendence* (max = 1). *Darwin's theory* and *Transcendence* concepts received a different scale because there was significantly more information in the instructional

text about the first concept compared to the latter. But the scale of the scores was not relevant for this study, because the analyses were conducted for each concept separately.

The rubric for Text 2 was composed of four concepts: *Contextualization of the text* (max = 2), *Process of strangulation* (max = 2), *Competition of the trees for reaching sunlight* (max = 2), and *Strategy of survival* (max = 2).

The rubric for Text 3 was composed of five concepts: *Debate* (max = 2), *Phonology* (max = 2), *Syntax* (max = 2), *Semantics* (max = 2), and *Symbol* (max = 2).

2.2.3. Computational resources

Gallito 2.0 (Jorge-Botana et al., 2013) software was used to implement both corpus training and the IR method. The initial LSA's semantic space was generated with a linguistic corpus of general knowledge that was extracted from digitalized texts from the Spanish Wikipedia. Specifically, a full list of the article titles of the Spanish Wikipedia was generated with a bot and a randomly selection of the contents of that list was made. Once the article titles were selected, then an automated bot extracted the digitalized texts of each article (the researchers supervised that the results were correctly processed). Then, the standard LSA procedure was applied to this sample of digitalized texts from the Spanish Wikipedia. The training corpus was composed of 404,436 documents and 39,566 unique terms. The log-entropy was used as the weighted function (Nakov et al., 2001), and a total of 300 dimensions were imposed for the latent semantic space. This semantic space was transformed later, using the IR method.

2.3.Procedure for human raters

Students were recruited, distributed between the three groups, and tasked to generate a constructed response (summary) of one of the three texts. A total of four human raters then evaluated the constructed responses of the students, using the rubrics described in *2.2. section*.

The instructional texts, the student constructed responses (summaries), and the human ratings can be found in the associated OSF project. Human raters assigned a score for each rubric concept in each summary, depending on the quality of the concepts within the summaries (inclusion of some sub-topics and a coherent discourse). As can be observed in the *2.2. section*, those scores ranged typically from 0 to 2 (omission of the concept vs. coherent and full explanation of the concept, respectively). Specifically, one of the human raters evaluated the constructed responses of all the texts, and another, the constructed responses of each separate text to achieve reliability for the assessments. Thus, two different human raters then evaluated the constructed responses of the students in each text. The final rubric score was established as the mean evaluation of both raters.

A total of 285 responses were collected, but 30 summaries were excluded from the study to maintain a reliable target concept evaluation due to low inter-rater reliability. In general, human raters gave similar scores for each concept in each summary, but when human raters scored concepts differentially (i.e., showing a difference of more than half of the total score) the summary was not included in the study. This exclusion criterion was established to assume that the target concept is clear in the evaluated summary because if no consensus is found for the target concept among the human raters in that summary, it would not be clear whether the computational methods should detect that concept in that specific summary or not. Once the human assessments were established and the sample was filtered, the rubric scores were compared with those generated by the IR and the PCS methods in different statistical tests (that is, human raters conducted a blind assessment of the student summaries).

2.4. Procedure for computational methods

The first step to generate both computational scores is to generate LSA's latent semantic space, reducing it to 300 dimensions with SVD. Then, the PCS and the IR methods

that were described in the *1.1. and 1.2. sections* were computed using the same latent semantic space.

In the PCS method, instructional texts must be segmented in order to generate different fragments that comprise the contents of each of the concepts to be evaluated. Thus, it is necessary to obtain as many fragments as concepts we want to evaluate. Then, a PCS score is obtained through the cosine between the vector representation of each summary and each instructional text fragment. See *Figure 1* as a hypothetical example of the PCS method. This has been a widely applied strategy to detect specific contents and we use it as a baseline.

In the IR method, some lexical descriptors per concept must be generated by human evaluators. Lexical descriptors are obtained by consensus searching for a good definition of each concept in the latent semantic space. The quality of the lexical descriptors of each concept is evaluated by analyzing the semantic neighborhood of their vector representation. Typically, three descriptors per concept are enough, as no improvement was found by adding a higher number (Martínez-Huertas et al., 2018). *Table 1* shows the lexical descriptors used for each concept in the present study. Semantic space was then transformed using these lexical descriptors (see Olmos et al., 2014). The original latent space was then transformed into a meaningful one, generating a semantic space whose vector coordinates represented specific semantic contents (those *p* concepts that were used to transform semantic space; see *Formula 1*). Following this, *p* concepts per text were represented in the semantic space and *p* scores were obtained (weighting by the *W* index; see *Formula 2*). Each *p* score represents the IR conceptual concept score. See *Figure 2* as a hypothetical example of the IR method.

Table 1. Lexical descriptors per text used to transform the latent semantic space.

	Concepts	Lexical descriptors
Text 1 (*Darwin's Theory of Evolution*)	Earth's age (C_1)	Hutton Buffon earth
	Lamarck (C_2)	Lamarck characteristics acquired
	Darwin's expedition (C_3)	Beagle Galapagos finches
	Darwin's theory (C_4)	selection natural evolution
	Transcendence (C_5)	polemic biology modern
Text 2 (*Strangler Trees*)	Contextualization of the text (C_1)	tree strangle Brasil
	Process of strangulation (C_2)	kill asphyxiation roots
	Competition of the trees for reaching sunlight (C_3)	competition lights sun
	Strategy of survival (C_4)	adaptation survival survive
Text 3 (*Theory of the Evolution of Language*)	Debate (C_1)	Evolution Neuroscience Paleontology
	Phonology (C_2)	Phonetics Articulation Deafness
	Syntax (C_3)	Syntax Sentence Macromutation
	Semantics (C_4)	Semantics Meaning Sign
	Symbol (C_5)	Symbol Abstraction Flexibility

Note. All lexical descriptors were translated from Spanish.

Once the student summaries were automatically evaluated, the computational scores of both the IR and the PCS methods were compared with those provided by the rubrics of the human raters.

3. RESULTS

Results have the following structure: (1) Human inter-rater reliability was calculated as the *Intraclass Correlation Coefficient* (ICC) (Shrout & Fleiss, 1979) to analyze the credibility of the external criteria. This measure reflects absolute agreements between measurement and ICCs above .75 are considered as indicators of good reliability (Koo & Li, 2016); (2) The predictions of the IR and PCS methods were compared in a topic-detection task analyzing the convergent and discriminant validity as standardized β coefficients from multiple linear regressions, to predict the human evaluations using the computational scores as covariates; (3) The multicollinearity of IR and PCS scores was analyzed in order to explain the differential performance of both methods in the topic-detection task; and (4) the semantic representations of the IR and PCS methods were compared in order to analyze the capacity of

the twos to endow their vector representations with meaning. All the statistical analyses were conducted with IBM SPSS Statistics 19.

3.1. Human inter-rater reliability

Inter-rater reliability was calculated as the ICC for each concept and the total rubric scores. In all the texts, a high reliability was obtained for the evaluation of the concepts. Also, a high reliability was obtained for total rubric scores (see *Table 2*). While these results set the credibility of rubric scores, their reliability was increased through the deletion of some of the summaries in the analysis.

Table 2. Inter-rater reliability for each concept in the assessment rubrics

	C_1	C_2	C_3	C_4	C_5	**Total score**
Text 1	.94	.91	.92	.86	.93	.94
Text 2	.93	.94	.70	.90	-	.94
Text 3	.93	.96	.93	.89	.86	.97

Note. Reliability measures were established through Intraclass Correlation Coefficients (all were statistically significant with $p<.01$). Text 1 = *Darwin's Theory of Evolution*. Text 2 = *Strangler Trees.* Text 3 = *Theory of the Evolution of Language*. C_1-C_5 = Concepts 1 to 5.

3.2. Predicting concepts through Inbuilt Rubric (IR) and Partial Contents Similarity (PCS) methods

In order to test the quality of the semantic representations of the IR and PCS methods, the relationship between the computational scores and the human rubric scores were exhaustively analyzed. Thus, different multiple linear regressions were conducted to predict each of the human evaluations of concepts from *Table 1*, using the computational scores of each method separately as covariates. Then, the multiple linear regression coefficients were used to analyze the convergent and the discriminant validity of these computational scores in predicting the quality of the evaluated concepts (see *Table 3*). Specifically, the standardized β coefficients were used to determine the most predictive computational score in each multiple

linear regression model. The following model was used to test the convergent and discriminant validity of the *k* computational scores:

$$C_{tkci} = \beta_{t11} * Score_{t11i} + \beta_{t22} * Score_{t22i} + \cdots + \beta_{tkc} * Score_{tkci} + \cdots + \beta_{tpc} * Score_{tpci} + e_{tkci} \quad [1]$$

where C_{tkci} was the human evaluation of the *k* computational scores (1, 2, …, *k*, …, *p*) of the *c* concepts (in our study, c ranges from 1 to 5 concepts) for subject *i* (1, 2, … , *i*, …, *n*) in each instructional text *t* (in our study, *t* ranges from 1 to 3 texts), $Score_{tkci}$ represents the *k* computational score of concept *c* in the instructional text *t* for subject *i*, and β_{t11}, β_{t22}, β_{tkc}, β_{tpc} was the predictive coefficients of each computational score *1, 2, ..., k, ..., p* for text *t* (β is an standardized regression coefficient and then the intercept was 0). Thus, an appropriate prediction (and then, convergent and discriminant validity) for concept *c* in text *t* occurs when β_{tkc} was the highest regression coefficient of computational score *k* for concept *c* in the equation for text *t.*

The IR method presented a strong convergent and discriminant validity for each semantic representation in evaluating the human assessments in Text 1 and Text 2. In both texts, the best predictor of each concept was always its own IR method concept. For Text 1, the highest standardized β coefficient values range from .29 to .70 (b_{144}=4.34, *SE*=1.88, *t*=2.31, *p*<.05, β_{144}=.29; and b_{111}=7.72, *SE*=.88, *t*=8.73, *p*<.01, β_{111}=.70; respectively). For Text 2, the highest standardized β coefficient values range from .40 to .67 (b_{244}=.07, *SE*=.02, *t*=3.94, *p*<.01, β_{244}=.40; and b_{222}=.07, *SE*=.01, *t*=8.52, *p*<.01, β_{222}=.67; respectively). In these texts, the best predictor of each concept was always its own IR method concept. In Text 3, the performance of the IR method was not as accurate as in the previous texts. Whilst there is considerable convergent and discriminant validity of each IR concept in evaluating human assessments as shown in the highest standardized β coefficient values ranging from .35 to .58 (b_{355}=4.36, *SE*=1.21, *t*=3.62, *p*<.01, β_{355}=.35; and b_{311}=6.79, *SE*=1.10, *t*=6.19, *p*<.01, β_{311}=.58;

respectively), the third and fourth concepts ("Syntax" and "Semantics") did not have the highest standardized β coefficient: .15 and .23 (b_{333}=1.81, *SE*=1.43, *t*=1.27, *n.s.*, β_{333}=.15; and b_{344}=2.98, *SE*=1.25, *t*=2.39, *p*<.05, β_{344}=.23; respectively) in comparison to the standardized β coefficients of dimensions three and four that were equal to .33 and .45 (b_{323}=3.51, *SE*=1.15, *t*=3.05, *p*<.01, β_{323}=.33; and b_{314}=4.68, *SE*=1.08, *t*=4.33, *p*<.01, β_{43}=.45; respectively). The results of Text 3 were different from those obtained in the other texts and some explanations for this are discussed below.

The PCS method presented less convergent and discriminant validity for each concept in evaluating the human assessments in all the texts, than the IR method. But it is noteworthy that the PCS scores presented a differential performance depending on the text. In this case, the best performance was in Text 1 where only the first three PCS scores were the best predictors of its own concepts: b_{111}=6.03, *SE*=.56, *t*=10.82, *p*<.01, β_{111}=.81; b_{122}=1.72, *SE*=.94, *t*=1.83, *n.s.*, β_{122}=.22; and b_{133}=4.30, *SE*=1.26, *t*=3.40, *p*<.01, β_{133}=.46. The fourth and fifth concepts were predicted by other PCS scores in Text 1: b_{134}=1.19, *SE*=.62, *t*=1.91, *n.s.*, β_{134}=.28; and b_{135}=1.77, *SE*=.64, *t*=2.78, *p*<.01, β_{135}=.37; respectively. In Text 2, two PCS scores were the best predictors of its own concepts: b_{222}=.42, *SE*=.80, *t*=.53, *n.s.*, β_{222}=.09; and b_{233}=.97, *SE*=.57, *t*=1.72, *n.s.*, β_{233}=.28. The first and fourth concepts were predicted by other PCS scores in Text 2: b_{231}=1.21, *SE*=.82, *t*=1.49, *n.s.*, β_{231}=.25; and b_{234}=1.84, *SE*=.79, *t*=2.33, *p*<.05, β_{234}=.36; respectively. The worst performance was found for Text 3, where none of the PCS scores was able to accurately predict concepts. Specifically, the following PCS scores were found to be the best predictors of concepts one to five, respectively: b_{341}=2.92, *SE*=1.39, *t*=2.09, *p*<.05, β_{341}=.32; b_{312}=2.30, *SE*=1.18, *t*=1.96, *n.s.*, β_{312}=.26; b_{313}=1.56, *SE*=.79, *t*=1.97, *n.s.*, β_{313}=.26; b_{314}=1.88, *SE*=.88, *t*=2.15, *p*<.05, β_{314}=.28; and b_{325}=.39, *SE*=.94, *t*=.41, *n.s.*, β_{325}=.06.

Table 3. Results from multiple linear regressions to detect concepts (C_1-C_5) using the Inbuilt Rubric (IR) method and Partial Contents Similarity (PCS) method scores.

			Est_1 (*se*)	β_1	Est_2 (*se*)	β_2	Est_3 (*se*)	β_3	Est_4 (*se*)	β_4	Est_5 (*se*)	β_5	R^2
Text 1	**IR**	C_1	7.72** (0.88)	.70	1.50 (2.00)	.07	0.90 (1.80)	.04	2.20* (1.03)	.18	3.89* (1.85)	.18	.51
		C_2	0.17 (0.81)	.02	4.34* (1.88)	.29	3.18 (1.69)	.21	0.19 (0.96)	.02	0.62 (1.73)	.04	.10
		C_3	3.40** (0.89)	.31	-1.19 (2.00)	-.06	10.78** (1.80)	.55	1.69 (1.03)	.15	5.29** (1.85)	.27	.43
		C_4	0.38 (0.47)	.08	1.86 (1.05)	.20	1.90* (0.95)	.21	1.53** (0.54)	.29	2.41* (0.97)	.25	.25
		C_5	0.11 (0.51)	.02	2.77* (1.16)	.26	2.19* (1.04)	.22	1.25* (0.60)	.21	3.02** (1.07)	.32	.28
	PCS	C_1	6.03** (0.56)	.81	-1.11* (0.88)	-.10	-0.57 (0.94)	-.06	1.54 (0.83)	.14	0.23 (0.61)	.03	.66
		C_2	0.21 (0.60)	.04	1.72 (0.94)	.22	1.48 (1.00)	.22	1.17 (0.89)	.16	-0.82 (0.65)	-.15	.18
		C_3	-0.02 (0.75)	.00	1.09 (1.19)	.11	4.30** (1.26)	.46	1.32 (1.21)	.13	-1.38 (0.82)	-.19	.28
		C_4	-0.43 (0.37)	-.14	-0.47 (0.58)	-.10	**1.19 (0.62)**	**.28**	0.84 (0.55)	.18	0.49 (0.40)	.15	.17
		C_5	-0.67 (0.38)	-.19	0.03 (0.60)	.01	**1.77** (0.64)**	**.37**	1.44* (0.56)	.37	0.28 (0.41)	.08	.31
Text 2	**IR**	C_1	0.14** (0.02)	.64	0.00 (0.01)	.04	-0.01 (0.02)	-.04	-0.01 (0.02)	-.04			.40
		C_2	0.08** (0.02)	.31	0.07** (0.01)	.67	0.00 (0.02)	.02	0.00 (0.02)	.00			.60
		C_3	0.06** (0.02)	.36	-0.00 (0.01)	-.06	0.06** (0.01)	.48	0.00 (0.01)	.03			.42
		C_4	0.09** (0.02)	.35	0.01 (0.01)	.08	0.02 (0.02)	.12	0.07** (0.02)	.40			.47
	PCS	C_1	-0.22 (0.54)	-.05	-0.30 (0.71)	-.07	**1.21 (0.82)**	**.25**	-1.14 (0.77)	-.28			.06
		C_2	-1.17 (0.61)	-.25	0.42 (0.80)	.09	-0.15 (0.92)	-.03	-0.57 (0.87)	-.12			.07
		C_3	0.31 (0.37)	.10	-0.60 (0.49)	-.19	0.97 (0.57)	.28	-1.20* (0.54)	-.40			.14
		C_4	0.84 (0.52)	.19	-1.23 (0.69)	-.27	**1.84* (0.79)**	**.36**	-2.29** (0.75)	-.52			.24
Text 3	**IR**	C_1	6.79** (1.10)	.58	3.85** (1.33)	.29	-2.07 (1.66)	-.13	2.45 (1.26)	.17	1.41 (1.54)	.08	.41
		C_2	4.39** (1.43)	.32	7.05** (1.74)	.45	-0.88 (2.16)	-.05	2.20 (1.65)	.13	1.68 (2.01)	.08	.26
		C_3	2.71** (0.95)	.30	**3.51** (1.15)**	**.33**	1.81 (1.43)	.15	2.11 (1.09)	.19	1.70 (1.33)	.12	.26
		C_4	**4.68** (1.08)**	**.45**	2.05 (1.31)	.17	1.80 (1.63)	.13	2.98* (1.25)	.23	2.77 (1.52)	.17	.28
		C_5	0.51 (0.86)	.06	2.38* (1.04)	.26	-1.82 (1.29)	-.17	3.33** (0.99)	.31	4.36** (1.21)	.35	.26
	PCS	C_1	1.37 (1.02)	.18	-1.31 (1.29)	-.14	-0.78 (1.02)	-.11	**2.92* (1.39)**	**.32**	-0.68 (0.79)	-.11	.09
		C_2	**2.30 (1.18)**	**.26**	0.46 (1.50)	.04	-1.31 (1.18)	-.16	1.51 (1.61)	.14	-0.26 (0.92)	-.04	.08
		C_3	**1.56 (0.79)**	**.26**	0.59 (1.01)	.08	-0.90 (0.79)	-.16	0.50 (1.09)	.07	-0.47 (0.62)	-.10	.06
		C_4	**1.88* (0.88)**	**.28**	-0.21 (1.11)	-.03	-0.66 (0.88)	-.10	2.19 (1.20)	.27	-0.48 (0.68)	-.09	.14
		C_5	-0.19 (0.74)	-.04	**0.39 (0.94)**	**.06**	-0.40 (0.74)	-.08	0.30 (1.02)	.05	0.01 (0.58)	.00	.01

Note. Est = Unstandardized regression coefficient; s*e* = Standard Error. β = Standardized regression coefficient. ** = p<.01. * = p<.05. Shading cells = Adequate predictions. In bold = Inadequate predictions (higher β coefficients for different concepts). Text 1 = *Darwin's Theory of Evolution* (N=88). Text 2 = *Strangler Trees* (N=76). Text 3 = *Theory of the Evolution of Language* (N=91). C_1-C_5 = Concepts 1 to 5.

Given that previous results showed a higher performance of IR in comparison to the PCS method (see *Table 3*), a total score for both methods was estimated for each text, to test the overall impact of the choice of the method. Here, the total score of the human assessments was predicted by the scores of each method separately and, later, by scores of both the methods conjointly (see *Table 4*). As can be observed in the results of the multiple linear regression, the performance of IR was significantly higher than that of the PCS method for Text 1 (IR method: b=.05, SE=.01, t=6.97, p<.01, β=.71; PCS method: b=.69, SE=.58, t=1.20, *n.s.*, β=.12), Text 2 (IR method: b=.48, SE=.04, t=11.65, p<.01, β=.80; PCS method: b=-.32, SE=.29, t=-1.11, *n.s.*, β=-.08), and Text 3 (IR method: b=.08, SE=.01, t=10.99, p<.01, β=.77; PCS method: b=.28, SE=.45, t=.63, *n.s.*, β=.04). All the R^2 of the models were large: R^2=.64 for Text 1, R^2=.67 for Text 2, and R^2=.60 for Text 3, indicating a large proportion of predicted variance of the human assessments. It is noteworthy that the sum of the scores of the PCS method only reached an adequate prediction for Text 1 (b=3.81, SE=.46, t=8.31, p<.01, β=.67) with a R^2 equal to .47. The PCS method was incapable to reach adequate predictions for Text 2 (b=-1.15, SE=.46, t=-2.48, p<.05, β=-.28) and Text 3 (b=1.43, SE=.67, t=2.14, p<.05, β=.22) obtaining a R^2 equal to .05 in both instructional texts. In fact, the inclusion of the PCS score did not improve the current prediction of the scores of IR method (ΔR^2 from .00 to .01). This can also be seen in the coefficients of the multiple linear regressions where the predictions of IR method were more accurate than the ones of the PCS method in Text 1 (b=.05, SE=.01, t=6.97, p<.01, β=.71; and b=.69, SE=.58, t=1.20, *n.s.*, β=.12; respectively), Text 2 (b=.08, SE=.01, t=10.99, p<.01, β=.77; and b=.28, SE=.45, t=.63, *n.s.*, β=.04; respectively), and Text 3 (b=.48, SE=.04, t=11.65, p<.01, β=.80; and b=-32, SE=.29, t=-1.11, *n.s.*, β=-.08; respectively). These results mean that the overall prediction of the IR method was enough to predict a considerable part of the variance of the human constructed response assessments (R^2 from .60 to .67) comparing to the PCS method (R^2 from .05 to .47).

Table 4. Results from simple and multiple linear regressions to predict the overall quality of the summaries using the Inbuilt Rubric (IR) method and Partial Contents Similarity (PCS) method scores.

	Method	Text 1		Text 2		Text 3	
		***Est* (*se*)**	***β***	***Est* (*se*)**	***β***	***Est* (*se*)**	***β***
Simple linear regressions	**PCS**	3.81** (0.46)	.47	-1.15* (0.46)	-.28	1.43* (0.67)	.22
	R^2	.47		.05		.05	
	IR	0.05** (0.01)	.64	0.49** (0.04)	.82	0.08** (0.01)	.78
	R^2	.64		.67		.60	
Multiple linear regression	**PCS**	0.69 (0.58)	.12	-0.32 (0.29)	-.08	0.28 (0.45)	.04
	IR	0.05** (0.01)	.71	0.48** (0.04)	.80	0.08** (0.01)	.77
	R^2	.65		.68		.60	

Note. *Est* = Unstandardized regression coefficient; s*e* = Standard Error. *β* = Standardized regression coefficient. ** = $p<.01$. * = $p<.05$. Text 1 = *Darwin's Theory of Evolution* (N=88). Text 2 = *Strangler Trees* (N=76). Text 3 = *Theory of the Evolution of Language* (N=91).

3.3. Analyzing the multicollinearity and the similarity between the semantic representations of Inbuilt Rubric (IR) and Partial Contents Similarity (PCS) methods

Given that both the methods presented a differential performance in the topic-detection task plus the nature of both computational methods, we hypothesized that multicollinearity and the (dis)similarity between their semantic representations could explain these findings. In order to analyze whether both the computational methods generate similar semantic representations, the predictions of each concept were compared using Pearson correlation coefficients (see shading cells from *Table 5*). In *Table 5*, it can be observed that the similarity of the semantic representations of both computational methods was low, especially in Texts 2 and 3: *r* ranging from -.20 [-.40–.00] to .00 [-.22–.23], and from .11 [-.11–.31] to .21 [-.01–.39], respectively. On the contrary, the similarity between both computational methods was larger in Text 1 with *r* ranging from .09 [-.12–.29] to .74 [.61–.83]. Complementing these results by the findings from *Table 3*, we can conclude that the concepts were differentially represented by both computational methods considering the similarity between both computational methods and the human assessments. Thus, it seems that these semantic vector representations present a differential performance when they are endowed with meaning in favor of IR method representations. In *Table 5*, it can also be

observed that the PCS method presented much more multicollinearity than the IR method in almost all its scores (see the correlation matrix between the scores of each computational method in each text, especially those values in bold). In general, the mean Pearson correlation coefficient between the vector representations of the concepts of the PCS method was .39 for Text 1 (*r* ranging from .22 [-.04–.40] to .52 [.34–.67]), .47 for Text 2 (*r* ranging from .19 [-.11–.45] to .73 [.55–.83]), and .45 for Text 3 (*r* ranging from .24 [-.04–.47] to .56 [.39–.83]). On the contrary, the mean Pearson correlation coefficient between the vector representations of the concepts of the IR method was -.03 for Text 1 (*r* ranging from -.34 [-.49–-.18] to .44 [.29–.58]), .13 for Text 2 (*r* ranging from .02 [-.24–.30] to .40 [.18–.58]), and -.02 for Text 3 (*r* ranging from -.26 [-.48–-.01] to .50 [.30–.66]). These results show that the higher accuracy of the predictions of the IR method coordinates can be associated to their higher orthogonality, comparing to the PCS method scores. It is noteworthy that the coordinates of the IR method were more related in Text 3, probably because of the semantic space was not able to discriminate between much-related concepts (such as “semantics” and “symbols”). Thus, in addition to the (dis)similarity of their semantic representations, it can be concluded that the IR method has a higher discriminant validity than the PCS method.

Table 5. Multicollinearity and similarity between the semantic representations of Inbuilt Rubric (IR) and Partial Contents Similarity (PCS) scores as Pearson correlation coefficients [95%IC].

			IR method				**PCS method**				
			C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Text 1	**IR**	C_1	.05 [-.17–.28]	-.01 [-.23–.17]	-.23* [-.44–-.02]	-.03 [-.28–.23]	.74** [.61–.83]	.06 [-.19–.29]	.31** [.09–.29]	.01 [-.19–.21]	-.07 [-.27–.15]
		C_2	1	-.34** [-.49–-.18]	.14 [-.08–.36]	.44** [.29–.58]	.14 [-.06–.33]	.09 [-.12–.29]	.27* [.04–.46]	.05 [-.21–.27]	.30** [.07–.52]
		C_3		1	-.24* [-.43–-.03]	-.06 [-.25–.14]	-.07 [-.28–.14]	.07 [-.13–.24]	.27* [.06–.45]	.05 [-.18–.24]	-.14 [-.36–.10]
		C_4			1	-.06 [-.16–.30]	-.13 [-.36–.18]	.17 [-.09–.38]	.08 [-.17–.34]	.38** [.23–.54]	.71** [.58–.81]
		C_5				1	.25* [.03–.48]	.20 [-.01–.40]	.32** [.12–.50]	-.01 [-.19–.19]	.44** [.29–.57]
	PCS	C_1					1	.22* [-.04–.40]	**.51** [.31–.66]**	.20 [.02–.38]	.23* [.02–.43]
		C_2						1	**.50** [.32–.65]**	.35** [.18–.51]	**.50** [.31–.64]**
		C_3							1	**.52** [.34–.67]**	.45** [.27–.62]
		C_4								1	.40** [.21–.58]
Text 2	**IR**	C_1	.12 [-.09–.34]	.10 [-.10–.31]	.35** [.13–.52]		-.01 [-.26–.23]	-.13 [-.35–.11]	.00 [-.25–.25]	-.22 [-.45–.01]	
		C_2	1	-.24* [-.48–.03]	.02 [-.24–.30]		-.14 [-.38–.09]	-.07 [-.27–.15]	-.19 [-.39–.02]	-.02 [-.24–.20]	
		C_3		1	.40** [.18–.58]		-.07 [-.27–.16]	-.23 [-.40–.02]	.00 [-.22–.23]	-.11 [-.34–.10]	
		C_4			1		.06 [-.14–.24]	-.26* [-.48–-.04]	-.05 [-.27–.18]	-.20 [-.40–.00]	
	PCS	C_1					1	.44** [.21–.67]	.19 [-.11–.45]	.27* [-.01–.53]	
		C_2						1	**.55** [.33–.70]**	**.66** [.50–.77]**	
		C_3							1	**.73** [.55–.83]**	
Text 3	**IR**	C_1	-.25* [-.42–.05]	-.46** [-.60–-.28]	-.11 [-.33–.13]	.14 [-.08–.35]	.21 [-.01–.39]	-.04 [-.27–.18]	-.14 [-.35–.05]	.15 [-.08–.35]	.17 [-.03–.34]
		C_2	1	**.50** [.30–-.66]**	.25* [.08–.43]	-.26* [-.48–-.01]	-.05 [-.27–.19]	.13 [-.09–.31]	.07 [-.13–.26]	-.04 [-.30–.20]	-.03 [-.31–.23]
		C_3		1	.29* [.08–.52]	-.18* [-.38–.04]	.05 [-.16–.26]	.05 [-.16–.26]	.11 [-.11–.31]	-.03 [-.24–.16]	-.09 [-.32–.14]
		C_4			1	-.08 [-.27–.14]	.08 [-.12–.27]	.16 [-.08–.38]	.13 [-.06–.32]	.13 [-.09–.33]	-.04 [-.25–.16]
		C_5				1	.11 [-.13–.31]	.04 [-.19–.26]	.08 [-.15–.32]	.21* [-.01–.40]	.13 [-.11–.32]
	PCS	C_1					1	.32** [.09–.53]	**.53** [.36–.66]**	**.50** [.32–.65]**	.37** [.16–.56]
		C_2						1	**.54** [.37–.67]**	**.56** [.39–.70]**	.35** [.10–.55]
		C_3							1	**.53** [.34–.69]**	.24** [-.04–.47]
		C_4								1	**.53** [.31–.70]**

Note. ** = p<.01. * = p<.05. 95%IC was computed by bootstrapping (1000 random samples). Text 1 = *Darwin's Theory of Evolution* (N=88). Text 2 = *Strangler Trees* (N=76). Text 3 = *Theory of the Evolution of Language* (N=91). In bold = Large multicollinearity (r>.50). Shading cells = Similarity between the semantic representations of both methods.

4. DISCUSSION

In the present study, the semantic representations of two computational methods were analyzed in order to enhance computerized constructed response assessments of student summaries. The first method, Partial Contents Similarity (PCS), implements topic-detection as the cosine-based similarity between a text vector and another text vector in the latent semantic space. The second method, Inbuilt Rubric (IR), implements topic-detection in each text as a combination of different vectors in a meaningful semantic space. Thus, both methods use different vector representations in their respective semantic spaces. Specifically, these methods were tested in a topic-detection task, that is, in the computerized assessment of specific contents of constructed response assessments (in this case, student summaries).

The findings of the present study showed a great performance of the IR method in comparison to the PCS method, which uses a cosine-based similarity measure. In other studies, the IR method has shown great performance using overall scores in Automated Summary Evaluation (Martínez-Huertas et al., 2018, 2019), but this is the first time that its coordinates have been tested to represent specific semantic contents. In the topic-detection task, the assessments of IR showed great capacity to represent specific concepts in its non-latent semantic space (its coordinates correctly endowed semantic space coordinates with meaning). Thus, the IR method demonstrated here that an LSA-based method can produce non-latent and meaningful coordinates. In fact, its performance was excellent in two of the three texts in terms of convergent and discriminant validity, the third text being one with highly complex concepts (such as *debate* or *symbol*). Similarly, the performance of the PCS method was low in terms of convergent and discriminant validity. We found two explanations for this differential performance: (1) the high multicollinearity of the PCS scores (i.e., this method was not able to differentiate between similar concepts), and (2) the qualitatively different meaning of the semantic representations of both the methods. Further, this proposal

can be perceived in the differential functioning of computational models between expository and narrative texts (where usually expository texts present higher computational performances; Wolfe, 2005; León et al., 2006) since the first ones have idiosyncratic characteristics in terms of content and context-dependent polysemous meanings.

Thus, the IR method was proposed as a procedure to properly detect semantic topics in constructed response assessments. The use of orthogonalized multi-vector semantic representations in the IR method avoids the referential circle problem that occurs in computational models, which affirms that terms have meaning only in relation to their similarity to other terms (de Vega et al., 2012). Usually, computational semantic representations are based on the comparison of strongly related inputs (i.e., a vector representation with the vector representations of its own gold-standard criteria) (Turney, 2006; Turney & Pantel, 2010; see a similar rationale in Jain et al., 2020), but LSA's IR method (Olmos et al., 2014) procedure generates *p* meaningful coordinates, thereby transforming abstract and latent space into a meaningful one, whose coordinates represent these *p* concepts. Due to its importance to computationally define meaning, an interesting application could emerge for the study of ontologies: IR method could transform the latent semantic space using the acceptations of a word to analyze how important are the terms for each acceptation or which are the most predominant acceptations in different texts. In this line, the IR method can accurately evaluate semantic concepts from texts without human references (see other interesting proposals in automatic text summarization like Rojas-Simón et al., 2021).

4.1.Enhancing psychoeducational computerized assessments of constructed responses

The present study was circumscribed to Automated Summary Evaluation, but the capacity of the IR method to detect specific semantic concepts has both theoretical and practical implications.

The first implication is related to the scope of theories about knowledge representation, where the LSA was proposed as a disembodied learning machine that can learn meaning from symbols alone (Landauer, 1999). In this way, computational semantic measures extracted from the latent semantic space can lead to an accurate measurement of different psychological constructs (e.g., Kjell et al., 2019) and their relations with other variables (e.g., Azmi et al., 2019; Corcoran & Cecchi, 2020; Lalata et al., 2019; Susnea et al., 2017). Thus, in addition to the hypothetical cognitive mechanism that learns semantics from repeated episodic experiences in a linguistic environment, the IR method could model the activation of concepts on the semantic network. Therefore, many expert and intelligent system applications that use similarity measures for topic-detection tasks could be enhanced.

In the case of intelligent tutoring systems, the current findings open a window of opportunity for using the IR method to provide individualized and specialized feedback enhancing assessment for learning. In this way, valid assessments would be facilitated using this comprehensive framework for psychoeducational evaluation, taking advantage of the promotion of learning and instruction by facilitating feedback and self-assessment (Jonsson & Svingby, 2007). Classical studies that analyzed the benefits of automatic feedback in summarizing competence have shown considerable variability in their effect sizes, but they tend to have significant positive effects on the performance of the students (Kintsch et al., 2000; Wade-Stein & Kintsch, 2004; Landauer et al., 2009; Mohamadi, 2018). These tools can be valuable for students to self-monitor their progress and to check their own improvements over time, although these tools may not be suitable for awarding a final grade. In this way, the IR method could improve knowledge assessment in online education (Jorge-Botana et al., 2015; see similar educational applications for offline applications like Bellino & Bascuñán, 2020, or Tulu et al., 2021) or other online communities because of using a valid and interpretable semantic representation of texts. Further applications can be made in assessing contents in social network data like Twitter or assessing cross-lingual readability by means of

topic-detection to estimate the level of difficulty of texts in different languages. Moreover, the IR method could improve tag-based frameworks in different contexts, such as summarization of transcribed videos or differentiating psychological stages in transcribed chats due to having explicit topics that are established *a priori* in the computational model.

The semantic space coordinates of IR were able to represent semantic concepts, but a differential performance was observed in texts. Some explanations can be considered to understand the differential performance of both the methods in the present study that are referred to the generalist corpus information (no specialized knowledge about these concepts was present). While current results are promising, future research should focus on the interaction between the corpus characteristics (e.g., generalist vs. specific linguistic corpus) with the performance of different computational methods. While a generalist corpus was able to represent semantic concepts using the IR method, future research should analyze the advantages that a more specific corpus could bring to these results. These specific corpora could be achieved using the lexical descriptors used to transform the latent semantic space with IR method in order to enrich its knowledge of the target concepts. Thus, the design of the corpus for specific level groups (different educational levels or competences; see, for example, Jorge-Botana et al., 2018) would thus bring new possibilities to these computational models since the semantic representations depend on prior corpus knowledge.

4.2. Conclusions

In this paper, we conducted a between-subjects study to compare the convergent and discriminant validity of two computerized assessment methods designed to detect semantic topics in constructed responses (IR and PCS methods). While both methods are distributional models of language and use the same LSA prior knowledge, they produce different semantic representations. Results of this study showed that the predictive capacity of the IR method seems to be larger than the one of the PCS method. Also, it was observed that their

differential performance was related to two different properties of these computational scores: (1) The computational scores of the PCS method presented larger multicollinearity than the ones of the IR method, and (2) the computational scores of both methods present a considerable (dis)similarity as their evaluations of the same concepts seems to be very different. Thus, we concluded that topic-detection in constructed response assessments can be enhanced by the IR method using meaningful multi-vector representations of constructed responses, in comparison to classic LSA measures as the ones of the PCS method. As it has been shown, the multi-vector representations of IR method have different advantages: (1) First, they can represent different concepts and contents of a text, simultaneously mapping a considerable variability of contents in constructed responses; (2) Second, usual procedures compare the similarity between different text vector representations and cannot evaluate the absence of concepts, but the IR method is able to detect the absence of concepts when the scores of its semantic dimensions are near zero; and (3) Third, most importantly, the coordinates of the IR method are the result of orthogonal dimensions that can avoid the multicollinearity of common cosine-based vector representations. Given that recent and interesting LSA applications are starting to interpret its semantic space coordinates with meaning using less-refined strategies like varimax rotations (Kundu et al., 2015), it is noteworthy that the IR multi-vector representations could be a very interesting alternative for enhancing these educational technology research proposals. To summarize, the IR method properly endowed LSA dimensions with meaning, using a topic-based strategy. Thus, its meaningful semantic vector representations could enhance current computer-assisted language learning applications.

5. REFERENCES

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

Azmi, A.M., Al-Jouie, M.F., & Hussain, M. (2019). AAEE–Automated evaluation of students' essays in Arabic language. *Information Processing & Management, 56*(5), 1736-1752. DOI:10.1016/j.ipm.2019.05.008.

Bellino, A., & Bascuñán, D. (2020). Design and Evaluation of WriteBetter: A Corpus-Based Writing Assistant. *IEEE Access, 8,* 70216-70233.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. DOI:10.1037/h0046016.

Corcoran, C.M., & Cecchi, G. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5*(8), 770-779. DOI:10.1016/j.bpsc.2020.06.004.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407. DOI:10.1002/(SICI)1097-4571(199009)41:6<391:AID-ASI1>3.0.CO;2-9.

De Deyne, S., Navarro, D.J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General, 145*(9), 1228-1254. DOI:10.1037/xge0000192.

de Vega, M., Glenberg, A., & Graesser, A.C. (2012). *Symbols and Embodiment: Debates on Meaning and Cognition.* Oxford: Oxford University Press.

Dessus, P., & Lemaire, B. (1999). Apex, un système d'aide à la préparation d'examens. *Sciences et Techniques éducatives, 6*(2), 409-415.

Dronen, N., Foltz, P.W., & Habermehl, K. (2015, March). *Effective sampling for large-scale automated writing evaluation systems*. Proceedings of the Second (2015) ACM Conference on Learning@Scale (pp.3-10). ACM. http://dx.doi.org/10.1145/2724660.2724661.

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*(1), 53-80. DOI:DH8F-QJWM-J457-FQVB.

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, USA.

Graesser, A.C., Hu, X., Nye, B.D., VanLehn, K., Kumar, R., Heffernan, C., ... & Andrasik, F. (2018). ElectronixTutor: an intelligent tutoring system with multiple learning resources for electronics. *International Journal of STEM Education, 5*(15), 1-21. DOI:10.1186/s40594-018-0110-y.

He, H., & Lin, J. (2016). Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 937-948). DOI:10.18653/v1/N16-1108.

Hewitt, J. & Manning, C. D. (2019, June). *A structural probe for finding syntax in word representations*. Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers) (pp.4129–4138).

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A.C., & McNamara, D.S. (2007). Strengths, limitations, and extensions of LSA. In T.K. Landauer, D.S., McNamara, S. Dennis, & W. Kintsch, *Handbook of Latent Semantic Analysis* (pp. 401-426). New Jersey: Routledge. DOI:10.4324/9780203936399.ch20.

Jain, S., Seeja, K. R., & Jindal, R. (2020). A New Methodology for Computing Semantic Relatedness: Modified Latent Semantic Analysis by Fuzzy Formal Concept Analysis. *Procedia Computer Science, 167,* 1102-1109. https://doi.org/10.1016/j.procs.2020.03.412.

Jones, M.N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J.R. Busemeyer, Z. Wang, J.T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Mathematical and Computational Psychology (*pp. 232-254*)*. New York: Oxford University Press. DOI:10.1111/bjop.12201.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. DOI:10.1016/j.edurev.2007.05.002.

Jorge-Botana, G., Luzón, J.M., Gómez-Veiga, I., & Martín-Cordero, J.I. (2015). Automated LSA assessment of summaries in distance education: some variables to be considered. *Journal of Educational Computing Research, 52*(3), 341-364. DOI:10.1177/0735633115571930.

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., & Luzón, J.M. (2018). Word maturity indices with latent semantic analysis: why, when, and where is Procrustes rotation applied? *Wiley Interdisciplinary Reviews: Cognitive Science, 9*(c1457), 1-16. DOI:10.1002/wcs.1457.

Jorge-Botana, G., Olmos, R., & Luzón, J.M. (2019). Could LSA become a "Bifactor" model? Towards a model with general and group factors. *Expert Systems with Applications, 131*, 71-80. DOI:10.1016/j.eswa.2019.04.055 .

Kaur A., & Sasi Kumar M. (2019) Performance Analysis of LSA for Descriptive Answer Assessment. In H. Saini, R. Sayal, A. Govardhan, & R. Buyya (Eds), *Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 74*. Springer, Singapore. DOI:10.1007/978-981-13-7082-3_8.

Kjell, O.N., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods, 24*(1), 92-115. DOI:10.1037/met0000191.

Klein, R., Kyrilov, A., & Tokman, M. (2011, June). *Automated Assessment of Short Free-Text Responses in Computer Science using Latent Semantic Analysis*. Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education (ITiCSE '11). DOI:10.1145/1999747.1999793.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary street: computer-guided summary writing. In T.K. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds.), *The Handbook of Latent Semantic Analysis* (pp. 263-277). New Jersey: Routledge. DOI:10.4324/9780203936399.ch14.

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

Koo, T.K., & Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163. DOI:10.1016/j.jcm.2016.02.012.

Kundu, A., Jain, V., Kumar, S., & Chandra, C. (2015). A journey from normative to behavioral operations in supply chain management: A review using Latent Semantic Analysis. *Expert Systems with Applications, 42*(2), 796-809. DOI:10.1016/j.eswa.2014.08.035.

Lalata, J.A.P., Gerardo, B., & Medina, R. (2019, August). A correlation analysis of the sentiment analysis scores and numerical ratings of the students in the faculty evaluation. *Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition* (pp.140-144).

Landauer, T.K. (1999). Latent Semantic Analysis (LSA), a disembodied learning machine, acquires human word meaning vicariously from language alone. *Behavioral and Brain Sciences, 22*(4), 624-625. DOI:10.1017/S0140525X99382145.

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*, 211–40. DOI:10.1037/0033-295X.104.2.211.

Landauer, T.K., Lochbaum, K.E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory into Practice, 48*(1), 44-52. DOI:10.1080/00405840802577593.

Landauer, T.K., McNamara, D.S., Dennis, S., & Kintsch, W. (2007). *The Handbook of Latent Semantic Analysis*. New Jersey: Routledge. DOI:10.4324/9780203936399.

LaVoie, N., Parker, J., Legree, P.J., Ardison, S., & Kilcullen, R.N. (2020). Using Latent Semantic Analysis to Score Short Answer Constructed Responses: Automated Scoring of the Consequences Test. *Educational and Psychological Measurement, 80*(2), 399-414. DOI:10.1177/0013164419860575.

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. DOI:10.3758/BF03193894.

Madnani, N., Burstein, J., Sabatini, J., & O'Reilly, T. (2013). *Automated scoring of a summary-writing task designed to measure reading comprehension*. Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp.163-168).

Magliano, J.P., & Graesser, A.C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods, 44*(3), 608-621. DOI:10.3758/s13428-012-0211-3.

Martín-Loeches, M. (2016). Origen y evolución del lenguaje humano: Una perspectiva neurocognitiva. Retrieved from http://www.atapuerca.org/ficha/ZE7D1307E-A298-9B9E-5CF101F70223C275/origen-y-evolucion-del-lenguaje-humano-una-perspectiva-neurocognitive.

Martínez-Huertas, J.A., Jastrzebska, O., Mencu, A., Moraleda, J., Olmos, R., & León, J.A. (2018). Analyzing two automatic assessment LSA´s methods (Golden Summary vs Inbuilt Rubric) in summaries extracted from expository texts. *Psicología Educativa, 24*(2), 85-92. DOI:10.5093/psed2048a9.

Martínez-Huertas, J.A., Jastrzebska, O., Olmos, R., & León, J.A. (2019). Automated Summary Evaluation with Inbuilt Rubric method: An alternative to constructed responses and multiple-choice tests assessments. *Assessment and Evaluation in Higher Education, 44*(7), 1029-1041. DOI:10.1080/02602938.2019.1570079.

McNamara, D.S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science, 3*(1), 3-17. DOI:10.1111/j.1756-8765.2010.01117.x.

Mirończuk, M.M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications, 106*, 36-54. DOI:10.1016/j.eswa.2018.03.058.

Mohamadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Studies in Educational Evaluation, 59*, 29-40. DOI:10.1016/j.stueduc.2018.02.003.

Nakov, P., Popova, A., & Mateev, P. (2001, September). *Weight Functions Impact on LSA Performance*. Paper presented at the EuroConference Recent Advances in Natural Language Processing (RANLP'01). Sophia, Bulgary.

Olmos, R., Jorge-Botana, G., León, J.A., & Escudero, I. (2014). Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis. *Discourse Processes, 51*(5-6), 494–510. DOI: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(3), 359-373. DOI:10.1016/j.ipm2015.12.002.

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

Reddy, Y.M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. DOI:10.1080/02602930902862859.

Rehder, B., Schreiner, M.E., Wolfe, B.W., Laham, D., Landauer, T.K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*(2-3), 337–354. DOI:10.1080/01638539809545031.

Rojas-Simón, J., Ledeneva, Y., & García-Hernández, R.A. (2021). Evaluation of text summaries without human references based on the linear optimization of content metrics using a genetic algorithm. *Expert Systems with Applications, 167*, 113827. DOI:10.1016/j.eswa.2020.113827.

Roll, I., Aleven, V., McLaren, B.M., & Koedinger, K.R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267-280. DOI:10.1016/j.learninstruc.2010.07.004.

Saha, S.K., & Rao, D. (2019). Development of a practical system for computerized evaluation of descriptive answers of middle school level students, *Interactive Learning Environments*, 1-19. DOI:10.1080/10494820.2019.1651743

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014, April). Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 373-374). ACM. DOI:10.1145/2567948.2577348.

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

Suleman, R.M. & Korkontzelos I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications, 165*, 114130. DOI:10.1016/j.eswa.2020.114130

Susnea, I., Pecheanu, E., Dumitriu, L., & Cocu, A. (2017, April). *Exploring the connection between the students' creativity and summary writing skills.* 2017 IEEE Global Engineering Education Conference (EDUCON) (pp.347-350).

Tulu, C.N., Ozkaya, O., & Orhan, U. (2021). Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM. *IEEE Access, 9*, 19270-19280.

Turney, P.D. (2006). Similarity of semantic relations. *Computational Linguistics, 32*(3), 379–416. DOI:10.1162/coli.2006.32.3.379.

Turney, P.D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221. DOI:10.1080/00461520.2011.611369.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333-362. DOI:10.1207/s1532690xci2203_3.

Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F., & Hao, H. (2015). Semantic clustering and convolutional neural network for short text categorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers) (Vol. 2, pp. 352-357). DOI:10.3115/v1/P15-2058.

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*(2), 359-364. DOI:10.1037/0278-7393.31.2.359.