Automated Summary Evaluation with Inbuilt Rubric method: An alternative to constructed responses and multiple-choice tests assessments

Assessment & Evaluation in Higher Education

https://doi.org/10.1080/02602938.2019.1570079

José Á. Martínez-Huertas (josea.martinez@uam.es). Department of Cognitive Psychology at the Universidad Autónoma de Madrid.

Olga Jastrzebska (olga.j@cop.es). Department of Cognitive Psychology at the Universidad Autónoma de Madrid.

Ricardo Olmos (ricardo.olmos@uam.es). Department of Social and Methodological Psychology at the Universidad Autónoma de Madrid.

Jose A. León (joseantonio.leon@uam.es). Department of Cognitive Psychology at the Universidad Autónoma de Madrid.

Corresponding author:
José Á. Martínez-Huertas
Department of Cognitive Psychology
Faculty of Psychology
Universidad Autónoma de Madrid
Email address: josea.martinez@uam.es

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 & Evaluation in Higher Education, 44*(7), 1029-1041. https://doi.org/10.1080/02602938.2019.1570079.

ABSTRACT

Automated Summary Evaluation is proposed as an alternative to rubrics and multiple-choice tests in knowledge assessment. Inbuilt rubric is a recent Latent Semantic Analysis (LSA) method that implements rubrics in an artificially-generated semantic space. It was compared with classical LSA's cosine-based methods assessing knowledge in a within-subjects design regarding two validation sources: a comparison with the results of rubric scores and multiple-choice tests, and the sensitivity of predicting the academic level of the test-taker. Results showed a higher reliability for inbuilt rubric (from Pearson correlation coefficient .81 to .49) over the classical LSA method (from .61 to .34) and a higher sensitivity using binary logistic regressions and effect sizes to predict academic level. It is concluded that inbuilt rubric has a qualitatively higher reliability and validity than classical LSA methods in a way that is complementary to models based on semantic networks. Thus, it is concluded that new Automated Summary Evaluation approaches such as the inbuilt rubric method can be practical in terms of reliability and efficiency, and, thus, they can offer an affordable and valuable form of knowledge assessment in different educational levels.

Keywords: Automated Summary Evalution; Inbuilt Rubric; rubrics; summaries; multiple-choice tests

Assessment in higher education is experiencing a shift from traditional knowledge testing toward measures of higher-order thought processes and competences rather than tests of factual knowledge and lower-level cognitive skills. This shift has led to a strong interest in new methods for assessing knowledge acquired from spoken or written materials. For example, educators have often used multiple-choice tests to evaluate comprehension. There are undoubtedly advantages to multiple choice testing, including speed of assessment, the possibility of evaluating many different aspects in a short time-frame, low cost, objective reliability measures (test-retest, Cronbach's α, etcetera) and relatively simple analysis of the psychometric properties of items (see, for example: Abad, Olea, Ponsoda, & García, 2011). This form of assessment has its limitations, however, as multiple-choice tests are based on recognition memory which can be more superficial than that demanded by responses based on recall (Shapiro & McNamara, 2000; Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2007), so the measure of learning does not necessarily reflect deep understanding of the text. Thus, multiple-choice tests cannot face the educational challenges of current formative assessments (Ashenafi, 2017).

Several studies have demonstrated the importance of summarizing to evaluate comprehension and learning as well as how summaries play a major role in research on text comprehension (for example: Wade-Stein & Kintsch, 2004; Hong, 2016; Sung, Liao, Chang, Chen, & Chang, 2016; Saddler, Asaro-Saddler, Moeyaert, & Ellis-Robinson, 2017; Stevens, Park, & Vaughn, 2018). However, it is often taken for granted that students learn to summarize as they move to higher academic levels, without any explicit attempts to teach summarizing skills (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005). For researchers such as van Dijk and Kintsch (1983), summarizing involves the capacity to generalize, synthesize and write coherently. It thus goes far beyond reading, since it implies profound comprehension of what is read, often incorporating previous knowledge and active processes such as inference-making. In their model of comprehension, summarizing is essential to understanding, since it involves extracting and

possibly elaborating on the main contents of what is read, while at the same time eliminating superficial details. Summarizing, then, involves establishing relationships among important concepts, and presenting them in a coherent, organized manner. The information must be restructured, further abstracting it from the content of the text, allowing easier access to factual and conceptual knowledge in memory. Summarizing texts allows students to build on classroom information more effectively than simply rereading a text and, thus, it enables instructors to evaluate the extent to which the material has been understood. Learning to summarize is a central aspect of the comprehension process, so that reliably evaluating a summary is key to knowing whether a student has a deep understanding of a text. Summaries provide a valuable tool for evaluating comprehension, which is why there have been efforts to develop automatic approaches for assessing them. These developments suppose a useful online self-assessment tool to improve specific abilites (for example: Seifert & Feliks, 2018) and can produce a high-quality feedback for students (Dawson et al., 2018).

The key to reliably assessing summaries is the use of rubrics. Effective rubrics describe assessment criteria, levels of performance and the weights of each criterion (Jonsson & Svingby, 2007; Reddy & Andrade, 2010; Dawson, 2017). It must be clear that assessments should be independent of who is scoring or other temporospatial characteristics of the assessment, since inter-rater and intra-rater reliability can be influenced by several factors (Jonsson & Svingby, 2007). Thus, Automatic Summary Evaluation (ASE) avoids many problems related to the reliability of subjective judgments.

Automatic Summary Evaluation (ASE) using LSA models

In the present study, we will address both reliability and validity of ASE methods in order to model rubrics. Also, we will compare the effectiveness of ASE methods with summarizing and multiple-choice testing across a variety of reading comprehension tasks of varying complexity and students at different academic levels. In developing rubrics, it is recommended that reliability can be improved, for example, by using analytical instead of holistic scoring, using benchmarks,

being topic-specific, increasing the rating scale, or training at least two raters to use the rubric (see Jonsson & Svingby, 2007; see also Reddy & Andrade, 2010; or Dawson, 2017). Concretely, some of the characteristics that are recommended to improve the reliability, such as using analytical scoring, being topic-specific or increasing the rating scale, are implemented in our ASE proposal. Computational assessments usually consist of the creation of statistical models that predict human assessments using text characteristics based on human criteria (Shermis, Burstein, Higgins, & Zechner, 2010). In ASE, it is common to assess summaries by evaluating syntax, semantic content, or a combination of natural language processing measures. One of the most successful procedures to uncover semantic content of texts is Latent Semantic Analysis (LSA).

According to the taxonomy of Jones, Willist & Denis (2015), LSA is a distributional model of semantic memory based on a hypothetical cognitive mechanism that learns semantics from repeated episodic experiences in a linguistic environment. From a more classical point of view, LSA is defined as a theory or method that extracts and represents meanings of words using statistical methods (like *Singular Value Decomposition*) that are applied to a large corpus to measure similarities among words and group of words (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). In this way, LSA has been widely applied in recent decades as a natural language processing method to extract meaning from text, and it has a robust background that strengthens this approach as a computational representation of semantic memory (McNamara, 2010).

Classical LSA methods for scoring constructed responses are based on the comparison of vector representations of texts prepared by students with vector representations of texts that are written by experts, as these texts are considered gold-standard criteria for coherent, consistent and complete discourse content (for example: Foltz, Laham, & Landauer, 1999; Klein, Kyrilov, & Tokman, 2011, June). As described in Landauer, Foltz and Laham (1998), semantic distance between both vector representations (that is, the vector representations of both the students and the experts) is used as a measure of text quality since summaries written by experts are

considered as quality criteria. Although this method has been proven to be satisfactory for most purposes, the golden summary method has some limitations, as it collects all the ideas into a single vector representation (Olmos, Jorge-Botana, León, & Escudero, 2014), and the expert texts could contain some bias towards subjects or participants, such as giving more importance to some subjects in the summary or not being completely impartial across participants because the similarity between vector representations could be influenced by other factors such as syntax (Kintsch et al., 2000). In order to have a baseline that represents these classical LSA methods, the golden summary method was selected to be compared to a new method that computationally implements rubrics (Olmos et al., 2014).

Inbuilt rubric is a new method that extends the usefulness of LSA. The idea behind this method was proposed originally by Hu, Cai, Wiemer-Hastings, Graesser, & McNamara (2007), and later implemented in Olmos et al. (2014). Hu et al. (2007) pointed out that the *latent* nature of LSA (that is, abstract in nature because the dimensions are not meaningful or do not have explicit interpretations) and the predominant use of the cosine to detect semantic similarities, do not use all the information contained in the semantic space. They proposed a two-step mathematical solution to represent the latent information in explicit and meaningful dimensions: (1) find a new base with meaningful dimensions and (2) transform the entire LSA latent space to the new base. This proposal gives psychological plausibility to the LSA model because of the vector representations of word meanings are now linked to meaningful dimensions. Then, inbuilt rubric method transforms the original latent semantic space into a meaningful one using a new algebraic basis based-on independent word vectors. Thus, the two main steps of inbuilt rubric are: (1) to incorporate the main topics of a text into a new algebraic basis, and (2) to transform the old latent semantic space into a new, meaningful one through this basis (see Olmos et al., 2014). While classical LSA methods use the original and latent semantic space, inbuilt rubric uses a new one in which the dimensions have explicit intepretations.

How can inbuilt rubric be used to assess a summary? As this method represents an extension of the use of dimensionality with which LSA extracts meanings, we can derive a natural way to do automatic assessment. Instead of using the cosine measure between text vectors to assess the quality of any one of them, the inbuilt rubric method posits the use of dimension scores to assess the summaries. This can be done in a simple way: each summary is the sum of its word vectors and it is automatically projected onto the new, meaningful semantic space. Each dimension score is a measure of its corresponding topic, and adding up all the *k* meaningful dimensions will result in an overall score of the summary quality. After a student summary is projected onto meaningful dimensions, we can see whether the summary expresses something similar to each meaningful dimension. Thus, with this method we can detect subtle contents that are or are not included in a student summary, which the cosine measure cannot.

Inbuilt rubric method simulates a human rubric by defining a new basis that represents the main topics of a text, transforming the basis of the old abstract space into a new, meaningful one (where the dimensions can be explicitly interpreted as the main topics), and projecting and using the dimension scores of the summary to obtain an overall assessment of its quality. This new method has demonstrated good performance for its additional detection of more specific knowledge contained in a text across different and heterogeneous student samples who read a variety of texts, thus showing its generalizability (Olmos, Jorge-Botana, Luzón, Cordero, & León, 2016; Martínez-Huertas, Jastrzebska, Mencu, Moraleda, Olmos, & León, 2018). As inbuilt rubric is a novel assessment method, it is necessary to validate its robustness. Recently, Bejar, Mislevy & Zhang (2016) underlined the importance of validity in the development of automatic evaluation methods given that validity should not be considered an isolated or efficiency-enhancing process but as the foundation for valid inferences and decisions made about student progress, and as a process completely uninfluenced by the response modality of the measure. Thus, response modality is a factor that affects the estimated validity (for example, multiple-choice vs summaries assessments). Rubric assessments are recommended in student constructed

responses because of their reliability (Jonsson & Svingby, 2007; Dawson, 2017). On the contrary, multiple-choice tests are used to sweep text contents, and they have high content validity that forms a considerable part of construct validity in educational evaluation because of content-relevance and representativeness.

To sum up, the aim of the present study is to measure inbuilt rubric reliability and validity as opposed to those of classical LSA cosine-based similarity methods (like golden summary) by comparing both automatic assessments with expert rubrics in scoring summaries and multiple-choice tests. Given that Baartman, Bastiaens, Kirschner, and van der Vleuten (2007) analyzed reliability and validity in competence assessments and concluded that both can be evaluated by using multi-modal assessment tools, the inbuilt rubric method is proposed here as a viable alternative to more traditional knowledge assessments. Also, given that many automatic assessments are created for specific academic levels, and that validity should prevail over reliability in computational assessments (Bejar et al., 2016), the capacity of inbuilt rubric to discriminate between academic levels is tested through binary logistic regressions and effect size measures.

METHOD

Participants

The sample was composed of 100 students. 44 were high school Spanish students (24 females, average age was 17) and 56 participants were undergraduate Spanish students from the Autonomous University of Madrid (48 females, average age was 20). All students read and summarized two different texts that were equivalent in their characteristics (both were expository texts with similar numbers of words and topics, but with different levels of abstraction in their contents), and answered a multiple-choice test about their contents.

Materials and measures

Texts: Two expository texts were used in the present study. One text was about Darwin's theory of evolution and was extracted from the book *Great Ideas of Science* of Isaac Asimov (1969). This text had a length of approximately 1,300 words. The other text was about the evolution of language (Martín-Loeches, 2016). This text had a length of approximately 900 words. Both texts were written in Spanish. A different sample of 86 students rated the relative difficulty of the texts, and 77% of them said that the language evolution text was more difficult than the Darwin text.

Rubrics: A rubric was created by three experts trained in summary assessment supervised by two university instructors for each text in order to assess the quantity of knowledge that was included in the summaries and to grade the comprehension of the students. Rubrics were elaborated based on the discussion of judges about the necessary contents that should to be included in an ideal summary of each text. Conceptual axes were established in an inductive way, that is, they were extracted from relevant contents of the summaries of the expert judges (rubrics were composed of the common information that should be incorporated into an ideal summary of the text).

The Darwin text rubric was composed of five concepts: "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).

The language evolution text rubric was also composed of five concepts: "Language evolution debate" (max = 2), "phonology" (max = 2), "syntax" (max = 2), "semantics" (max = 2), and "symbol" (max = 2).

To obtain a reliable measure, the criteria from Jonsson & Svingby (2007) and León, Olmos, Escudero, Cañas, & Salmerón (2006) were used to create rubrics designed to assess summaries (both knowledge quantity and quality were assessed). Each conceptual axis evaluation considered both presence of some sub-topics and discourse coherence of those contents. These measures were used as reliable criteria to compare with the LSA's

assessments (Intraclass Correlation Coefficients -rubric scores reliabilities- ranged from .81 to .91, see *Results*).

Multiple-choice tests: An 18-item multiple-choice test was created for each text to assess the knowledge acquired by students with objective criteria. In order to preserve unidimensionality, a latent variable was calculated to obtain the total score of each student. This measure was used, like both rubrics, as a reliable criterion to compare with the computational assessments. These multiple-choice tests were created in order to evaluate the concepts that were extracted from ideal summaries in order to be equivalent measures with the rubric scores. Each conceptual axis was derived from five to six items that were created to asses acquired knowledge from the text. Then, a final 18-items version was established using a qualitative inter-rater assessment in order to preserve the content validity and considering an equal distribution of implicit and explicit items.

Corpus: A general domain corpus extracted from 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, September). A total of 300 dimensions were imposed for the latent semantic space.

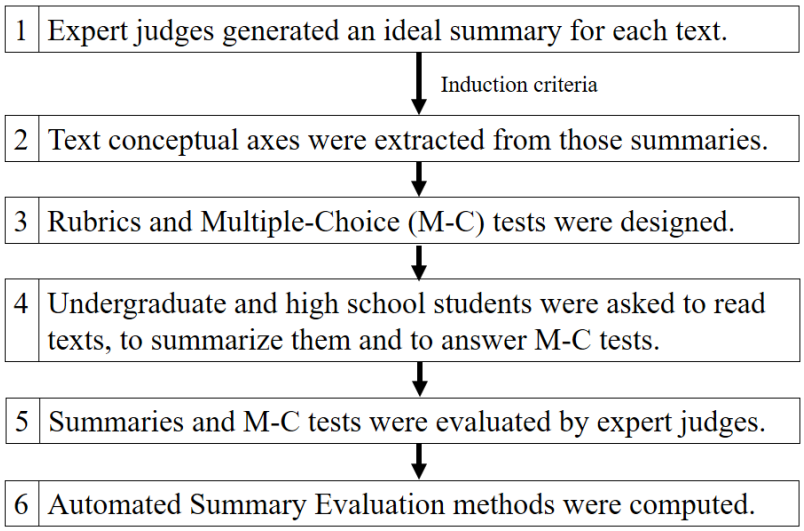
Software: Gallito 2.0 (Jorge-Botana, Olmos, & Barroso, 2013, July) was used to train and to establish the change of basis with the re-orthogonalization of the semantic space. This software makes it possible to perform the entire inbuilt rubric process method.

Procedure

In a within-subjects experimental design, the inbuilt rubric method was compared with the golden summary method as a baseline for the responses to the two texts. Both methods were compared with two external comprehension measures (rubric scores and multiple-choice tests). Possible effects of text content were controlled using a Balanced Incomplete Block Design (AB/BA) dividing the sample into two different groups in which the instructional texts were presented in a

different order. The principal interest of the results was to analyze Pearson correlation coefficient between LSA's assessments and external measures (rubric and multiple-choice test scores) in order to study the similarity between the automatic and external measures, and to evaluate the sensitivity to academic level for both automatic methods. A graphical representation of the procedure of the present study can be observed in Figure 1.

FIGURE 1. Diagram of the procedure of the present study.



Note. Six consecutive steps were followed in order to guarantee the reliability and validity of the results of the present study (including blind assessment).

First, three experts (who had been trained in summary assessment) independently generated an ideal summary for each text and were supervised by two instructors. In this way, three summaries of about 250 words for the Darwin text and for the language evolution text were created to extract their conceptual axes. Consensus criteria were established by induction (that is, they were extracted from relevant contents that were present in all the ideal summaries to find the common information that should be incorporated into a perfect summary of the text). Later, these summaries were used as the input for the golden summary method (used as a baseline), so it could be compared with the inbuilt rubric method.

Once the rubrics and the multiple-choice tests were designed, the sample was recruited. A group of undergraduate university students and a group of high school students were asked to read one of the texts and to create summaries of about 250 words from it, and subsequently, to answer the multiple-choice test. The same procedure was followed with the other text (reading, summarizing and test answering). Three expert judges assessed the summaries of each text on a 0 to 10 scale using the rubric described in *Materials and measures*. This assessment was established by a rubric that contained the conceptual dimensions of the text (five conceptual axes for each text). The assessment of the student summaries and the multiple-choice tests analyses were done by expert judges before any LSA assessment were realized (blind assessment).

Then, the Automated Summary Evaluation methods were computed. The golden summary method was calculated using the ideal summaries of the expert judges. This method transforms student summaries into vectors and extracts their cosines against the vector of the summary of the expert judge in order to measure the semantic similarities between both texts (Landauer, Foltz, & Laham, 1998; Foltz, Laham, & Landauer, 1999). A final golden summary score was considered per student as the mean of the cosine of his/her summary with each of the ideal summaries of the expert judges. In other words, the mean of the similarity of the summary of the student with all the summaries of the expert judges was used to obtain a reliable golden summary measure.

The inbuilt rubric method was calculated by capturing the meaning of the conceptual axis of the texts transforming the latent semantic space into a new semantic space (see Hu et al., 2007 or Olmos et al., 2014 for a complete description of this method). To transform a latent semantic space into a meaningful space it is necessary to establish some lexical descriptors (see Table 1 for the Darwin text descriptors and Table 2 for the language evolution text descriptors) that are supposed to capture each the conceptual axis of the text by projecting them into a vector space. Here, a lexical descriptor is a word or a group of words that represent a conceptual axis of the text in the latent semantic space. In the present study, three lexical descriptors were selected because, in another study, a higher number of descriptors did not improve the performance of the method

(Martínez-Huertas et al., 2018). Each summary was projected onto the semantic space, and the scores in each of the conceptual axes (that is, the scores in the transformed and explicit meaningful dimensions) were added to obtain a total score. Thus, the final score was calculated as the sum of the meaningful inbuilt rubric scores (those dimensions that transform latent semantic space).

TABLE 1: Lexical descriptors per dimension (conceptual axis) in the *Darwin text*

Conceptual Axis	Descriptors		
Earth's age	Hutton	Buffon	earth
Lamarck	Lamarck	characteristics	acquired
Darwin's expedition	Beagle	Galapagos	finches
Darwin's theory	selection	natural	evolution
Transcendence of the theory	polemic	biology	modern

Note: Each conceptual axis was composed of three descriptors that were originally written in Spanish.

TABLE 2: Lexical descriptors per dimension (conceptual axis) in the *Language evolution* text

Conceptual Axis	Descriptors		
Language evolution debate	Evolution	Neuroscience	Paleontology
Phonology	Phonetics	Articulation	Deafness
Syntax	Syntax	Sentence	Macromutation
Semantics	Semantics	Meaning	Sign
Symbol	Symbol	Abstraction	Flexibility

Note: Each conceptual axis was composed of three descriptors that were originally written in Spanish.

RESULTS

The results are presented in the following sequence: (1) Rubric scores and multiple-choice test reliabilities were calculated in order to assess the validity of these measures. (2) The automatic method reliabilities were presented as Pearson correlation coefficients between the automatic assessments and the results of both the rubric assessment and the multiple-choice tests, and hypotheses were tested using the likehood ratio. (3) Finally, the sensitivity to academic level was tested for both automatic methods.

Rubric scores and multiple-choice test reliabilities

With respect to the multiple-choice tests of the Darwin instructional text, a confirmatory factor analysis (CFA) was conducted in order to study the unidimensionality of the test. It discarded two out of the 18 items which did not contribute to the internal consistency ($Cronbach's\ \alpha = .66$). The unidimensional model fit reasonably well to the remaining 16 items ($\chi 2(149) = 174.1, p = .078;\ RMSEA = .04, 90\%\ CI\ [.00 - .07];\ CFI = .91;\ TLI = .90$). The inferred latent variable (adjusted for measurement error) was used as the total score.

A CFA was also conucted to study the unidimensionality of multiple-choice tests of Language evolution instructional text. It discarded three out of 18 items which did not contribute to the internal consistency ($Cronbach's\ \alpha = .70$). The unidimensional model fit reasonably well with the remaining 15 items ($\chi 2(132) = 161.0, p = 0.044;\ RMSEA = .05, 90\% CI\ [.01 - .07];\ CFI = .90; TLI = .88$). Again, the inferred latent variable was used as the total score.

To gain understanding of the validity of the expert rubrics and their reliabilities, the *Intraclass Correlation Coefficient* (ICC; Shrout & Fleiss, 1979) was calculated between the three raters in the Darwin text (N = 100 summaries), and a value of .814 was obtained. The same procedure was applied to the language evolution text, so the ICC obtained a value of .909. In Table 3 can be observed the rubric reliabilites as measured by Pearson correlation coefficient between expert judges for both texts. Both expert rubrics had a high reliability in this sample.

TABLE 3. Expert rubric reliabilities (as the Pearson correlation coefficients between expert judges).

Darwin text		EJ no. 2	EJ no. 3	EJ's mean
	EJ no. 1	.90	.82	.95
	EJ no. 2		.84	.96
	EJ no. 3			.94
Language evolution text		EJ no. 2	EJ no. 3	EJ's mean
	EJ no. 1	.96	.91	.98
	EJ no. 2		.90	.98
	EJ no. 3			.96

Note: EJ = Expert Judge. N = 100. All Pearson correlation coefficients were significant at p<.01 (bilateral)

Testing the reliability of automatic methods

An overall analysis was conducted in order to analyze whether there were differences in the reliabilities (as Pearson correlation coefficients) between the different assessments (that is, ASE methods, rubrics and multiple-choice tests). A likelihood ratio test (Raykov & Marcoulides, 2008, pp. 430) was used to test the null hypothesis of no text differences in the rubric-LSA reliabilities and in the multiple-choice test-LSA reliabilities.

Referring to the Darwin text results shown in Table 4, the rubric scores of the expert judges were correlated with the golden summary method ($r = .611$) and with the inbuilt rubric method ($r = .809$). A likehood ratio test showed significant differences between both correlations ($\Delta\chi2(1) = 18.9;\ p < .001$). Also, the inferred latent variable of the multiple-choice test was correlated with the rubric scores ($r = .744$), with the golden summary method ($r = .590$) and with the inbuilt rubric method ($r = .751$). According to the likelihood ratio test, there was a significant difference between the latter two correlations ($\Delta\chi2(1) = 10.68;\ p = .001$).

Referring to the language evolution text results shown in Table 4, the rubric scores were correlated with the golden summary method ($r = .596$) and with the inbuilt rubric method ($r = .775$). A likehood ratio test showed significant differences between both correlations ($\Delta\chi2(1) = 8.394;\ p = .004$). The inferred latent variable was correlated with rubric scores ($r = .703$), with the golden summary method ($r = .343$) and with the inbuilt rubric method ($r = .487$). According to the likelihood ratio test, there was not a significant difference between the latter two correlations ($\Delta\chi2(1) = 3.66;\ p = .0557$), but only a marginal difference.

TABLE 4. LSA methods' reliabilities using rubric scores and multiple-choice test results as the human criteria

		Multiple-Choice Test*	Golden summary method	Inbuilt rubric method
Darwin text	Expert Rubric scores	.74	.61	.81
	Multiple-choice Test*		.59	.75
	Golden summary method			.73
Language evolution text	Expert Rubric scores	.70	.60	.78
	Multiple-choice Test*		.34	.49
	Golden summary method			.53

All Pearson correlation coefficients were significant at p<.01 (bilateral). * = A latent variable was used to calculate the correlations (see *Rubric scores and multiple-choice test reliabilities*)

Sensitivity to academic level of automatic methods

To study the sensitivity to academic level of both automatic methods, different analyses were conducted: (1) a binary logistic regression first compared inbuilt rubric and golden summary (using academic level, high school and university, as the dependent variable), and (2) a descriptive comparison based on effect sizes (Hedges's g) was presented.

In the Darwin text, a binary logistic regression was conducted in order to compare inbuilt rubric and golden summary performances for the academic level classification. An inital binary logistic regression using inbuilt rubric and golden summary methods showed an improvement of correct classifications from 56% to 81% ($Nagelkerke's\ R^2 = .504$). Regression coefficients for inbuilt rubric showed that it is a good predictor ($p < .01$), while the golden summary did not reach statistical significance ($p = .12$). A second binary logistic regression was conducted with inbuilt rubric (to discount golden summary effects) reaching an improvement of correct classifications from 56% to 79% ($Nagelkerke's\ R^2 = .483$).

The same procedure was applied for the language evolution text. A first binary logistic regression was conducted using both methods as independent variables and resulted in an improvement of correct classifications from 55.6% to 70.7% ($Nagelkerke's\ R^2 = .295$). Regression coefficients for inbuilt rubric showed that it is a good predictor ($p < .01$), while

golden summary did not reach statistical significance ($p = .07$). A second binary logistic regression was conducted with inbuilt rubric (to discount golden summary effects) reaching an improvement of correct classifications from 55.6% to 73.7% ($Nagelkerke's\ R^2 = .257$).

An independent-samples t-test was applied ($p < .001$) to test if both groups had the same mean in all conditions (ASE methods, rubric scores, and multiple-choice tests). To compare the differences between academic levels in the same metric, Hedges's g was used as a less-biased version of Cohen's *d* (McGrath & Meyer, 2006). In the Darwin text, the rubric scores obtained a value of .938 while the multiple-choice test obtained a value of 1.455. In addition to these values, the inbuilt rubric method obtained a value equal to 1.488 and golden summary a value of 1.246. In the Darwin text, the LSA methods were more capable of discriminating between both groups than the rubric scores, as the inbuilt rubric was more accurate than the golden summary method. The same procedure was followed for the language evolution text. For expert rubric scores, a value of $g = 1.44$ was calculated, while the text comprehension test obtained a value of 1.678. The inbuilt rubric method obtained a value of $g = .95$ and golden summary a value of .728. In this case, the human measures had higher sensitivities than the LSA methods but, again, the inbuilt rubric was more accurate than the golden summary method.

DISCUSSION

Measuring comprehension for students who read a text, view a video, or listen to a lecture is of great importance to the design of educational materials. Comprehension has most commonly been measured by tests admisnistered shortly after the presentation of the materials. Due to the importance of summarization to reflect a deep understanding of a text, evaluating the performance of the students with summaries has a great number of advantages, but multiple-choice tests are characterized by their economical use of resources. For these reasons, an alternative to classical knowledge assessments has been proposed from advocates of Automatic Summary Evaluation (ASE). Originally, Hu et al. (2007) proposed a mathematical solution to

deduce measures of explicit meaning from an abstract and latent semantic space. This proposal strengthens and extends the usefulness and psychological plausibility of the LSA model. Olmos et al. (2014) implemented inbuilt rubric method and different empirical evidence showed its validity (Olmos et al., 2016; Martínez-Huertas et al., 2018), but there is still a strong necessity of probing its robustness. Following Bejar et al. (2016) who claimed that computational assessment methods need to search for valid inferences, the automatic evaluations of these modern and classical LSA methods were compared to rubric and multiple-choice test scores, and their sensitivities to academic level were tested through binary logistic regressions. As a human rubric that permits both detection and scoring of the main topics in a summary, inbuilt rubric is an LSA method that creates a new, meaningful semantic space enabling a human behavior-based rubric to capture and score the main topics in a summary or a text.

The reliabilities of LSA methods have reflected their good performance levels when they have been compared with rubric scores and multiple-choice tests. However, inbuilt rubric obtained higher similarities to rubric scores and multiple-choice tests than did the golden summary method. In this way, differences between methods in the Darwin and language evolution texts can be explained by the quality of the semantic representations of the corpus. That is, more specific concepts like “Darwin” or “Lamarck” are easily discriminated in a general corpus (as in Spanish Wikipedia) while concepts like “Semantics” or “Symbol” are abstract and very similar (being poorly differentiated even by humans). As Landauer and Dumais (1997) or Landauer, McNamara, Dennis, & Kintsch (2007) have affirmed, all LSA assessments depend on corpus characteristics, and the present study is no exception. These similarities between human and LSA performances can be considered evidence of the validity of automatic methods because the final objective is to implement a computational model that can emulate aspects of human semantic memory. Far from simple co-occurrence of words in a large corpus of texts, inbuilt rubric method can obtain useful results without the time-consuming retention of thousands of components using *Singular Value Decomposition*. This new method reduces semantic space into

300 dimensions in which some of them (*k*) are meaningful (they capture *k* concepts), thereby transforming latent space to a manageable size with the aim of assessing those *k* concepts.

Sensitivity analysis of both LSA methods showed good performance for both texts when comparing two academic levels (that is, presenting statistical differences and large effect sizes). Results showed that in the Darwin text, ASE methods showed a higher capacity to differentiate between academic levels than did the human judges (while for the language evolution text, the result was the opposite). A binary logistic regression was conducted to classify students into high school or university levels using their automatic assessments. When inbuilt rubric and golden summary methods were compared to differentiate groups of students, inbuilt rubric captured more variance and was the only method that reached statistical significance for both texts. Thus, new methods like inbuilt rubric have a higher sensitivity to secondary variables that can have great importance (academic level) for the comprehension of students. Future research should try to analyze the capacity of the method to classify users in relevant variables such as intelligence, motivation, personality, or, even, clinical status. ASE approaches, like the inbuilt rubric method, can have a high level of performance when classifying students and emulating human assessments due to its objective measures, adding the advantages of both rubrics and multiple-choice tests.

Although the groups belonged to high school and university communities, the mean age difference was 3 years (that is, high school students were ending their educational period in order to access university). Even when both groups presented an adequate competence in comprehension and summarization, inbuilt rubric method was sensible to their differences. Usually, it is considered that the comprehension of the students is more superficial when it is assessed with multiple-choice tests rather than by student-constructed responses (Shapiro & McNamara, 2000), but this was not a problem in the present study because students were prepared to make a summary from the text, and, only when they had completed it, to answer a

multiple-choice test. Possibly, a different experimental design could find differences in human assessment comparisons if not all students had to do both tasks.

LSA's ASE methods (specially the inbuilt rubric method) showed a great similarity with human measures (reliability and validity) and considerable sensitivity to academic level in binary logistic regression results (validity). In their professional practice, many professors and teachers have a great amount of work that could be complemented and helped using these automated assessments to evaluate their students in an ecologically valid way (that is, as if a teacher or professor would be evaluating them using a rubric and providing them feedback). For example, the inbuilt rubric method was very satisfactory for a university sample of 864 students when used to improve text comprehension and writing skills (Olmos et al., 2016). The present study shows how inbuilt rubric can maintain rubric benefits, like maintaining reliability and validity and promoting learning and instruction of students (Jonsson & Svingby, 2007). Inbuilt rubric can also aid in evaluating knowledge and skills through multiple-vector representations in semantic space and, thus, having the capacity to provide feedback for each conceptual axis. Other benefits of using rubrics, in educational performance, are assessment transparency, anxiety reduction, promotion of feedback processes, self-efficacy improvement, and student self-regulation support (Panadero & Jonsson, 2013).

The cognitive demands of multiple-choice tests are related to information recognition without necessarily requiring deep understanding of the text. Exposing students to summarization tasks, on the other hand, offers a means of evaluating deeper understanding of the material due to the necessity of recalling data and explanations, thus demanding comprehension of texts. The shortage of resources increases the use of multiple-choice tests and does not allow teachers to provide feedback individually (Wade-Stein & Kintsch, 2004), although there are some tools that could let students know their own progress through feedback about their ability levels in a longitudinal way. Rubrics are a solution to such problems of assessment and individual feedback because of being formative and summative approaches to assessment (Jonsson & Svingby, 2007;

Dawson, 2017). Inbuilt rubric method is an ASE tool that implements an analytical and topic-specific rubric to evaluate concepts included in texts with the capacity to provide high-quality feedback. With this method, the improvement of the agreement between expert judges using rubrics is not necessary, and the election of lexical descriptors to transform the semantic space seems to be a good alternative as a practical procedure for knowledge assessment.

Since classical LSA's cosine-based similarity (Landauer & Dumais, 1997) has been widely applied and tested in different contexts and tasks, it has obtained substantial validation that has encouraged the advance of other semantic models (Jones et al., 2015). Following this trajectory, new LSA methods like inbuilt rubric (Olmos et al., 2014) should be applied in different contexts and tasks because they seem to perform better in semantic content-detection tasks involving active semantic networks (in this case, assessing comprehension and knowledge from summaries). While Baartman et al. (2007) proposed that reliability and validity of competence evaluation should be obtained with multi-modal assessment programs, Bejar et al. (2016) declared that validity should be the focus of the development of automatic assessments. As in the present study, this strategy should be extended to other methods with the aim of improving educational measurement. Bejar et al. (2016) affirmed that every aspect of assessment design is relevant to test scoring and that it must be practical in terms of reliability and efficiency, so that automated scoring methods can offer an affordable and valuable form of knowledge assessment.

ACKNOWLEDGEMENTS

We are thankful to Professor James Juola for his insightful comments and support.

DECLARATION OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

REFERENCES

Abad, F.J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en Ciencias Sociales y de la Salud* [Measurement in Social and Health Sciences]. Madrid: Síntesis.

Ashenafi, M.M. (2017). Peer-assessment in higher education–twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education, 42*(2), 226-251. https://doi.org/10.1080/02602938.2015.1100711.

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

Baartman, L.K., Bastiaens, T.J., Kirschner, P.A., & van der Vleuten, C.P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*(2), 114-129. https://dx.doi.org/10.1016/j.edurev.2007.06.001.

Bejar, I.I., Mislevy, R.J., & Zhang, M. (2016). Automated Scoring with Validity in Mind. In A.A. Rupp & J.P. Leighton (Eds.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 226-246). Oxford: Wiley Blackwell. https://dx.doi.org/10.1002/9781118956588.ch10.

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education, 42*(3), 347-360. https://doi.org/10.1080/02602938.2015.1111294.

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2018). What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education.* Advance online publication. https://doi.org/10.1080/02602938.2018.1467877.

Foltz, P.W., Laham, D., & Landauer, T.K. (1999). Automated Essay Scoring: Applications to Educational Technology. In B. Collis & R. Oliver (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA '99)* (pp. 939-944). Seattle, USA: Association for the Advancement of Computing in Education (AACE).

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33*(1), 53-80. https://dx.doi.org/10.2190/DH8F-QJWM-J457-FQVB.

Hong, W. (2016). The Effect of Summarizing Task and Interaction on Korean Middle School Students' Reading Comprehension. *Studies in English Education, 21*(1), 39-71.

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. https://dx.doi.org/10.4324/9780203936399.ch20.

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. https://dx.doi.org/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. https://dx.doi.org/10.1016/j.edurev.2007.05.002.

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.

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*(2), 87-109. https://dx.doi.org/10.1076/1049-4820(200008)8:2;1-B;FT087.

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). https://dx.doi.org/10.1145/1999747.1999793.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284. https://dx.doi.org/10.1080/01638539809545028.

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

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. https://dx.doi.org/10.1037/0033-295X.104.2.211.

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://dx.doi.org/10.3758/BF03193894.

Martín-Loeches, M. (2016). *Origen y Evolución del Lenguaje Humano: Una Perspectiva Neurocognitiva* [Origin and Evolution of Human Language: A Neurocognitive Perspective]. 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. http://doi.org/10.5093/psed2048a9.

McGrath, R.E., & Meyer, G.J. (2006). When effect sizes disagree: the case of r and d. *Psychological Methods, 11*(4), 386. https://dx.doi.org/10.1037/1082-989X.11.4.386.

McNamara, D.S. (2010). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science, 3*(1), 3-17. https://dx.doi.org/10.1111/j.1756-8765.2010.01117.x.

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., & McNamara, D.S. (2007). Assessing and improving comprehension with latent semantic analysis. In T.K. Landauer, D.S., McNamara, S. Dennis, & W. Kintsch, *Handbook of Latent Semantic Analysis* (pp. 207-225). New Jersey: Routledge. https://dx.doi.org/10.4324/9780203936399.ch11.

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. https://dx.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*(3), 359-373. https://dx.doi.org/10.1016/j.ipm2015.12.002.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9,* 129-144. https://dx.doi.org/10.1016/j.edurev.2013.01.002.

Raykov, T., & Marcoulides, G.A. (2008). *An Introduction to Applied Multivariate Analysis.* Routledge: New York. https://dx.doi.org/10.1111/j.1751-5823.2009.00074_18.x.

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

Saddler, B., Asaro-Saddler, K., Moeyaert, M., & Ellis-Robinson, T. (2017). Effects of a summarizing strategy on written summaries of children with emotional and behavioral disorders. *Remedial and Special Education, 38*(2), 87-97. https://doi.org/10.1177/0741932516669051.

Seifert, T., & Feliks, O. (2018). Online self-assessment and peer-assessment as a tool to enhance student-teachers' assessment skills. *Assessment & Evaluation in Higher Education*. Advance online publication. https://doi.org/10.1080/02602938.2018.1487023

Shapiro, A.M., & McNamara, D.S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research, 22*, 1–36. https://dx.doi.org/10.2190/M811-G475-WKMX-X0JH

Shermis, M.D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N.S. Petersen (Eds.), *International Encyclopedia of Education, Vol.4 (3rd ed.)* (pp. 20-26). Oxford: Elsevier. https://dx.doi.org/10.1016/B978-0-08-044894-7.00233-5.

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

Stevens, E.A., Park, S., & Vaughn, S. (2018). A review of summarizing and main idea interventions for struggling readers in Grades 3 through 12: 1978–2016. *Remedial and Special Education*. Advance online publication. https://doi.org/10.1177/0741932517749940.

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

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

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