Journal of Research in Applied Linguistics
ISSN: 2345-3303 – E-ISSN: 2588-3887 – https://rals.scu.ac.ir

Please cite this paper as follows:

Vázquez-Cano, E., Ramírez-Hurtado, J. M., & Díez-Arcón, P. (2024). Teacher vs. machine correction: Comparing assessments of students' reading comprehension and writing skills. *Journal of Research in Applied Linguistics*, *15*(2), 6-21. https://doi.org/10.22055/rals.2024.45498.3186

Research Paper

Teacher vs. Machine Correction: Comparing Assessments of Students' Reading Comprehension and Writing Skills

Esteban Vázquez-Cano[1], José M. Ramírez-Hurtado[2], & Paz Díez-Arcón[3]

[1]Corresponding author, Department of Didactics and School Organization, Faculty of Education, Universidad Nacional de Educación a Distancia, Madrid, Spain; *evazquez@edu.uned.es*

[2]Department of Economics, Quantitative Methods and Economy History, Faculty of Business, Pablo de Olavide University, Sevilla, Spain; *jmramhur@upo.es*

[3]Department of Foreign Languages and their Linguistics, Faculty of Philology, Universidad Nacional de Educación a Distancia, Madrid, Spain; *pdiez@flog.uned.es*

Received: 07/12/2023 Accepted: 25/09/2024

Abstract

This article presents research that compared two correction techniques applied to a PISA text summary question written by 30 Spanish students aged 14-16, one by automatic correction software (G-Rubric) and the other by 30 Spanish language teachers varying in age, sex, and classroom experience. The methodology was a parametric approach based on latent class analysis using Latent Gold 4.5 software, and correspondence analysis. In the results, the Euclidean distances between each individual and the system were measured as low, medium or high dissimilarity, based how close the teachers' assessment was to that of the correction software. The results showed a first cluster, comprised of teachers whose correction scores exhibited a significant correlation with the tool, represented the quartile of younger and less experienced teachers. This stands in contrast to a second cluster, characterized by "high" dissimilarity, which consisted of older and more experienced teachers whose corrections deviated notably from the system, yielding scores lower than those produced by the tool.

Keywords: Automatic Correction; Human Corrector; Reading Comprehension; Assessment.

1. Introduction

The use of automatic correction software to assess text comprehension is becoming more frequent, and continuous technological advances in the assessment of writing and text comprehension make it a potentially effective tool for teachers in secondary education. Automatic correction systems can cut time spent correcting students' work to allow teachers to dedicate more effort to helping students improve their reading comprehension skills and written production; similarly, automatic feedback can help students produce better written work and sharpen their understanding of texts. Assessing a student's understanding of a text is subjective in that it involves variables that cannot always be controlled or extrapolated to other contexts. On the other hand, many automatic correction programs fail to appreciate sufficiently the creative and analytical skills that students display in their writing (Usener, Gruttman, Majchrzak & Kuchen, 2010). The viability and efficacy of this software vary according to language and complexity.

For example, most studies in this field have focused on automatic correction software for English, an analytical language of concise syntax and little derivation, whereas there are few automatic assessment tools for written work and reading comprehension available in Spanish, with the consequent lack of research into their viability and efficacy in Spain. This research is based on the following assumption: "The automated assessment of open-ended responses by G-Rubric can be equivalent to that performed by a real teacher." For this reason, this investigation set out to analyze the efficacy, and differences and similarities in the assessments of style and content in a text summary written in Spanish by Spanish students generated by one of the most promising automatic correction programs, G-Rubric. To this end, we will

formulate the following research questions: (1) What degree of similarity exists between the G-Rubric correction and the one that a real teacher would perform? (2) Are there significant differences or similarities between the G-Rubric correction and the one performed by a teacher according to his/her age and years of experience? The intention was to test the viability and applicability of such software in evaluating students' levels of reading comprehension and written work in Spanish.

2. Literature Review

2.1. Automatic Assessment of Students' Written Production

There is a range of terms used to define machine-based assessment, such as e-Assessment, Computer-Based Assessment (CBA), Computer-Assisted/Aided Assessment (CAA), computerized testing and computer-administered testing (JISC, 2007; Redecker, 2013). Likewise, automatic correction of written work has been described as Automated Essay Grading (AEG), Automated Essay Evaluation (AEE), Computer Essay Grading, Computer-Assisted Writing Assessment, Automated Essay Scoring (AES), Automated Writing Evaluation (AWE), Computer-based Assessment Methods" (CbAS), Automatic Essay Assessment, Automatic Essay Evaluation, and Computer-based Essay Marking System (CBEM) (Ericsson & Haswell, 2006; Landauer, 2003; Shermis & Burstein, 2003; Warschauer & Ware, 2006; Zhang, 2013). Automated Writing Evaluation (AWE) provides automated feedback, and its function is more related to assistance in writing. Such tools (computer-generated feedback) help students in their writing by correcting spelling, style and grammar, and are included in this definition as they can be used to help teachers assess students' work (Barret, 2015). In contrast, Automated Essay Scoring (AES) gives automated scores based on Mathematical models for the organizational, syntactic and mechanical aspects of writing (Ware, 2011).

These systems regularly use NLP techniques (E-Rater-Criterion e Intellimetric-MyAccess); Latent Semantic Analysis (LSA) (Intelligent Essay Assessor (IEA); Jess; MarkIT; G-Rubric); and, more discreetly, Bayesian methods (BETSY) (Dikli, 2006). NLP uses statistical methods to train the system based on large quantities of linguistic data extracted from real texts. LSA processes legible texts using a linguistic corpus and represents the words that appear in sentences, paragraphs, or essays by statistical calculation. These are statistical models of the use of vocabulary and can make semantic comparisons between two texts (Foltz, 1996). The advantage of LSA is its capacity to imitate the choice of vocabulary made, and the human judgement involved, by focusing on the content at the semantic level. To be able to assess written expression, LSA needs to be trained in specific domains of knowledge to determine the conceptual relevance of a text by comparison with others in the same field.

All these systems center on the ability of computer technology to assess and classify written work (Shermis & Bustein, 2003) and are a considerable advance on automatic multiple-choice systems, among others (Pérez-Marín, Alfonseca Cubero & Rodríguez Marín, 2006). According to the literature, style and content are the most important elements to be assessed in students' written output, by either holistic or rubric-based assessment. Csapó, Ainley, Bennett, Latour and Law (2012) stated that automatic evaluation first needs to itemize the components that merit points in order to then assess them and score the work. To do so, these programs must compute a set of variables like structure, the complexity of the words used and their distribution in the text, etc., as a combination that determines the score assigned to a text in a similar way to how a teacher might do so (Ben-Simon & Bennet, 2007). According to the former, technological tools for automated text correction are an interesting resource for teachers and students at the different stages of education, despite the challenges and obstacles. This technology has recently been applied in test assessments as an educational and diagnostic approach (Bennett, 2010; Bridgeman, 2009; Chen & Cheng, 2008; Csapó et al., 2012; Eggen & Verschoor, 2006), in which analysis was made of rhetorical and formal aspects of writing, such as lexis, syntax, discursive and grammatical structures, word selection and content development (Chen & Cheng, 2008). However, this type of technology-supported assessment often falls short when called on to assess competences, critical thought, complex problem solving, decision taking and students' creativity in the development of self-regulated learning

2.2. Correction of Open Questions and Summaries

When correcting students' work and exams, the more conscientious teachers use a range of strategies to get a clearer picture of a student's progress: correction by sets of questions rather than a full exam, limiting examination time to avoid fatigue, double checking their assessments, using self- and peer- assessment, providing formative and corrective feedback, or using correction rubrics, with the aim of achieving greater objectivity (Fernández-Alonso, Woitschach &

Muñiz Fernández., 2019; Hashemian & Farhang-Ju, 2018; Mirzaee & Tazik, 2014; Soleimani & Rahmanian, 2014; Vázquez-Cano et al., 2021). In extreme cases, awareness of this problem can lead to replacing this type of "subjective" assessment for "objective" multiple choice tests, although many authors have stated that this type of test "cannot provide a thorough evaluation as to whether the student has understood the concepts presented in class".

A key question is to what extent teachers allow automatic software to correct students' writing and text comprehension, and how to improve the reliability and validity of these assessment systems. The results in the literature are mixed as inferred from the previous literature review; some studies conclude that human and automated correction systems are equally reliable, while others state that automatic systems offer more credible results; other investigations emphasize the drawbacks in both. With software continually improving in sophistication, investigators are working to make automatic systems more reliable in the assessment of written work in the form of open answers and text summaries (Barrada, Olea, Ponsoda & Abad, 2006; Blumenstein, Green, Fogelman, Nguyen & Muthukkumarasamy, 2008; He, Hui & Quan, 2009; Noorbehbahani & Kardan, 2011). According to Powers, Burstein, Chodorow, Fowles and Kukich (2002), any improvements in assessment software need to focus on identifying repetition in questionable logic in written work. Furthermore, Rudner and Gagne (2001) stated that automatic correction tools are fast, reduce costs and avoid the potential human failings of fatigue and inconsistency. Similarly, Zhang (2013) concurred in the superiority of automatic correction systems as they are unaffected by external factors, and their speed at assessing grammar, lexis, style, text organization and development (Intelligent Essay Assessor and E.rater) compared to human assessment. Along the same lines, Bejar (2011) stated that the advantage of these tools lies in their consistency, which supersedes potential human shortcomings and instability in the cognitive processes, as well as their ability to assess written constructions that might escape human scrutiny.

Automatic correction software's usefulness is perceived by the quality of feedback it provides on the formal aspects of writing rather than content development (Chen & Cheng, 2008), which could hinder the acquisition of the "deep knowledge" students need to express themselves correctly and confidently in their writing (Benítez & Lancho, 2016; Zhang, 2013; Vázquez-Cano et al., 2023), and to augment their mental competences (Perelman, 2012). That said, there are results in the literature that show how this type of instrument, in particular Intelligent Essay Assessor (IEA), can help develop students' critical thinking (Wohlpart, Lindsey & Rademacher, 2008). Yet, in the study by Power et al. (2002), one participant stated that E-rater was incapable of appreciating aspects like individuality, humor or poetic inspiration, which not only dampens students' creative capacity but also inclines them to adopt the style of composition validated by the system, giving rise to potential manipulation (Chen & Cheng, 2008; Zhang, 2013), although the authors emphasize that is not easy to achieve.

Literature also highlights outstanding limitations of these systems. Zhang (2013) insisted that technology alone was insufficient for assessing students' writing and text comprehension skills; automated feedback on grammatical and textual aspects is weak and requires a teacher's judgement to provide clarification, explanation and guidelines for any feedback to be useful to the student. Additionally, human correctors can meet specific students' needs regarding written feedback such as those related to the "argument, logical order, transition, clarity, and references decision" (Hoomanfrad, Jafarigohar, Jalilifar, & Masum (2018, p.24). Powers et al. (2002) showed how programs such as E-rater were incapable of making accurate assessments without teacher intervention. These authors defended the use of Automated Essay Evaluation (AEE) tools as an initial assessment of students' work, to be followed by teacher review. Pérez-Marín et al. (2006) suggested that this order should be reversed, with the software validating the teacher's assessment. However, if the aim is to find and validate automatic correction software that can truly understand written competence and grade it accordingly, that wish is still far from being satisfied, since.

Regarding validation, Shermis and Hammer (2013) acknowledged the reliability of scores generated by nine AEE instruments when compared to those of human correctors and Klobucar, Elliot, Deess, Rudniy and Joshi (2013) concluded that Criterion and its E-Rater system were reliable as corrector tools in a similar comparison. Additionally, Attali and Burstein (2006), and Powers, Burstein, Chodorow, Fowles and Kukich (2000) showed how a second version of E-Rater could be a viable alternative when measuring written competence, as the user can see the relevance that the software designates to each points-based category and understand the tool's decision processes, in comparison to human correctors' assessments. On the other hand, Perelman (2013) cast doubt on the methodology deployed to reach such conclusions, and Ericsson and Haswell (2006) questioned the validity of industry-backed studies for the commercial

implications. Likewise, Bridgeman et al. (2012) generalized that there were positive comparable links between human and automated correction methods. LanguageTool was studied by researchers to establish its reliability against human correctors, and strong correlations were found between grammar correction and mechanical errors (Crossley, Bradfield & Bustamante, 2019). Finally, Bennet (2010) and Bridgeman (2009) discovered a high degree of similarity to human correction, although the results corresponded to open-response tests and were not very extensive. However, Bennet (2010) stated that automated correction tools such as IEA, Intellimetric, Project Essay Grader and E-Rater had been used to satisfactory effect in assessing long texts.

Evidently, the results are inconclusive and further studies are needed to explore the differences and similarities that exist between human and machine correction. For example, Warschauer and Ware (2006) synthesized correlation trials between Automated Essay Scoring (AES) and humans, as well as humans vs. humans, across a range of tools to assess written production (Intellimetric; E-rater; Critique*;* IEA), concluding that in 95% of cases the scores were similar between both groups, which supported the reliability of this technology in performing assessments. In contrast, the ACARA NASOP research team (2015), and Rudner et al. (2005), found significant differences between the scoring by Intellimentric and IEA, and that by human correctors. Wolhpart et al. (2008) found a concordance rate of 54% among human correctors compared to 81% by Automated Essay Evaluation (AEE). Bridgeman et al. (2012) judged that just as aspects of task correction can differ between human and automated correctors, the same is also true among human correctors of different demographics and training.

2.2. Automatic Assessment Systems of Written Work in Spanish

Although there are automatic correction tools specifically designed for Spanish, there are no studies that measure these tools' reliability when tested against human correctors. Most automatic correction systems are made for the English-speaking market (da Cunha, 2020), so research on such tools in other languages is rare (Amorim & Veloso, 2017). Nevertheless LanguageTool, an "open-source code system for multiple languages" (da Cunha, 2020: 46), has been adapted for use in Spanish to correct spelling in the written work of students of Spanish as a second language (Blázquez & Fan, 2019). The Stilus system can proof-read texts in Spanish to highlight grammatical, spelling, semantic and stylistic errors (Villena et al., 2002). Another system for Spanish language is Correctme, which was developed by Spain's National Distance Education University (UNED), can detect and correct spelling and grammatical errors in texts written by native Spanish students with a statistical analysis of word frequency, word pairs, or bigrams, based on a corpus of Spanish texts (San Mateo, 2016). Finally, da Cunha (2020) described Linguakit as a tool that can analyze texts to detect spelling, lexis, grammar and style errors in the Galician language, but which can be adapted for use in other languages, according to its creators.

G-Rubric (https://pre.psicoee.uned.es/grubric/), which is the instrument analyzed in this paper, is an automatic free writing assessment software program based on Latent Semantic Analysis (LSA) and designed by researchers at UNED's Department of Evolutionary Psychology and Education (Guillermo de Jorge Botana, José María Luzón, Ricardo Olmos Albacete and Alejandro Barroso). Currently, G-Rubric belongs to Semantia Lab, a technology-based startup that has the academic and institutional support of the Universidad Nacional de Educación a Distancia (UNED). This system can assess students' responses to text interpretation and written comprehension exercises in Spanish, in the dimensions of style (writing) and function (content) (Martín-Monje & Barcena, 2024), which can provide personalised feedback using an open technology (Díez-Arcón & Martín-Monje, 2021). This is novel in that there are currently no tools with these features available for Spanish. LSA, and other similar forms of text semantic analysis, have been successfully integrated into automatic programs for assessing written work (ACARA NASOP research team, 2015; Attali & Burstein, 2006).

This software applies a 1–10-point scale to correct content and assess appropriateness, as teachers do, and provides an overall assessment, like teachers, of the quality and suitability of the written work (composition and style). It functions according to a linguistic corpus specific to a thematic area and is based on the concept of spatial vector models that use linear algebra to assign lexical items in an n-dimensional vector space. The corpus is processed and expressed as a matrix that includes its terms and paragraphs. The next step is to assign a weighted entropy to each term to predict relevant asymmetries in lexical frequency. The weighting indicates its degree of focalization in a thematic area, that is, how specific the term is. Thus, heavily weighted terms are specific and limited to certain contexts, in relation to other equally specific terms, while lightly weighted terms are general. The advantage of LSA is that it applies the Singular Value Decomposition (SVD) dimension reduction technique to this matrix, which means it can identify each term or

paragraph using relatively few dimensions (about 300) and discards the rest as "noise" in the common use of language. SVD reveals semantic relations not indicated by the knowledge areas of the corpus (Landauer et al., 1998), so more apparently unrelated knowledge can be inferred.

3. Methodology

This research is based on the following hypothesis: "The automated assessment of open-ended responses by G-Rubric can be equivalent to that performed by a real teacher." For this, the main aim of this research is to describe the differences and similarities in the correction of text comprehension exercises and written work by Spanish students, between a computerized automated system and human correctors. To this end, we will formulate the following research questions: (1) What degree of similarity exists between the G-Rubric correction and the one that a real teacher would perform? (2) Are there significant differences or similarities between the G-Rubric correction and the one performed by a teacher according to his/her age and years of experience?

The study recruited 30 Spanish students aged 14-16 (16 females, 14 males) at five Spanish secondary school centers selected randomly; these were given a reading comprehension test from PISA 2009 (https://acortar.link/3n01n). The text included an explanation at the beginning and spanned 385 words, and there was also an explanatory introduction. The students had to make a summary of no less than 60 words and no more than 100 words. Once the summaries were completed, photocopies of each summary were made and sent to the teachers for correction. First, each of the 30 teachers corrected 30 summaries and G-Rubric program corrected the 30 summaries according to style and content. Based on those items, the system establishes the following classifications: requires considerable improvement (0-80), requires some improvement (81-90), acceptable (90-100). Content is scored from 0-10 (Grading between 0 and 10 points, as would be given by a teacher, on the quality and conceptual correctness of the content). While style is marked on a scale of 0-100; this score is determined by analyses for spelling, syntax (degree of complexity in the construction of simple, subordinate and coordinated sentences to form and relate ideas) and a general review of written expression according to: (1) "Coherence and cohesion": Evaluate the flow and logical connection between the ideas presented in the summary. (2) "Accuracy and relevance": Determine whether the summary accurately captures the key points of the original text and omits irrelevant details. (3) "Clarity and conciseness": Evaluate the student's ability to clearly and concisely communicate the essential ideas of the original text. (4) "Proper use of structure and format": Check whether the summary follows an appropriate structure, such as introduction, development, and conclusion and (5) "Originality and own voice": Consider whether the student has been able to express the ideas of the original text in his/her own words and in his/her own style.

As an example, Figure 1 presents the program's response to one student's summary which, according to its correction system, scored low (Figure 1).

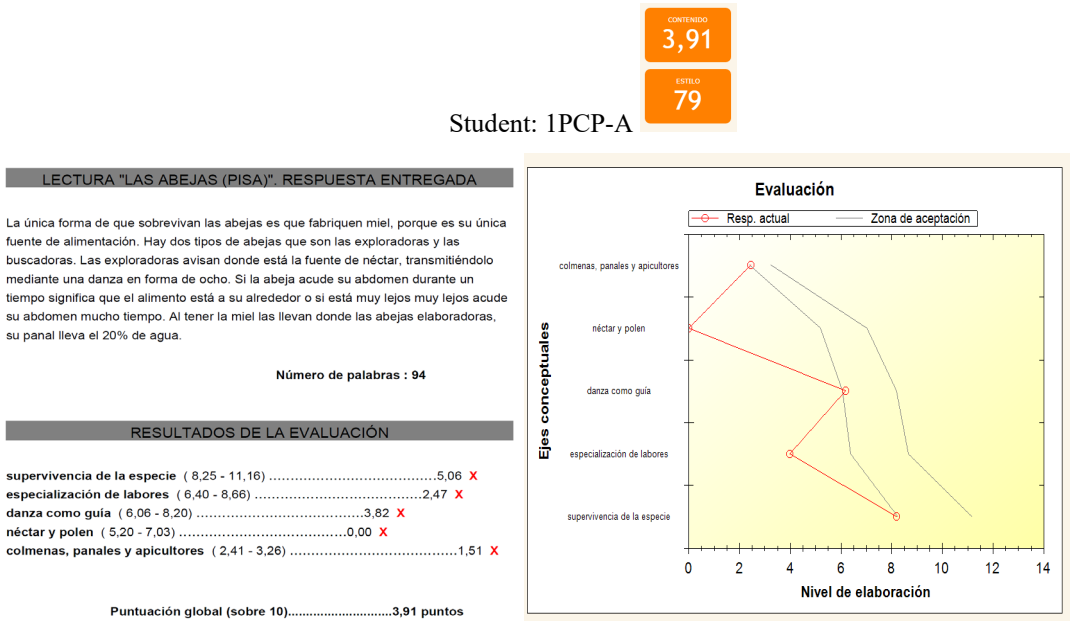

Figure 1. *An Example of G-Rubric Software Correction*

Following G-Rubric's correction of the summaries, 30 Spanish language instructors from five different areas of Sapin -Madrid, Castilla la Mancha, Andalucía, Extremadura, and Asturias- corrected the original texts. The instructors were chosen at random. Out of the 51 corrections that were submitted back, 21 were disapproved (5 due to the instructor's failure to provide informed consent for study participation, and an additional 16 to ensure a homogeneous sample reflecting the study variables of sex, age, and years of teaching experience). The teachers corrected the text on paper, with the only condition that they should mark the texts according to their normal class criteria; anonymity was guaranteed, and they were not informed that their corrections and scoring would be compared to other teachers' or to that of automatic correction software. The aim was to ensure that the factors conditioning correction were minimal. Upon receiving the revised texts, the educators were prompted to provide their express consent to be included in the study. They were also briefed about the procedure for analysis, which involved comparing their anonymously scored texts to those of other educators and the G-Rubric program, as well as calculating the three variables of age, sex, and years of teaching experience.

To identify the teacher profile that most closely matched the G-Rubric correction software's assessment, the Euclidean distance was calculated from the values that each individual apportioned to all the styles and contents, and those by the correction system. Given that there were 30 styles and 30 contents, the Euclidean distance between each individual and system was calculated as follows:

$$d(x,y) = \left(\sum_{k=1}^{60}((y(k) - x(k))^2\right)^{1/2}$$

x being any individual and y representing the system.

The Euclidean distances for each teacher and the system are expressed in ranges of dissimilarity, such that the greater the distance between the teacher and the system's scores, the greater the discrepancy between both. To define these three categories, the three quartiles of the Euclidean distances were found, and each was classified thus:

- Dissimilarity < Q1? "Low dissimilarity"
- Q1? Dissimilarity? Q3? "Medium dissimilarity"
- Dissimilarity > Q3? "High dissimilarity

Latent Class Analysis (Lazarsfeld, 1950; Lazarsfeld & Henry, 1968; Bartholomew et al. 2002; Vermunt & Magidson, 2002, 2003, 2005) was used to identify potential teacher profiles in their corrected texts using Latent Gold 4.5 (Statistical Innovations). LCA is a parametric model that uses the data obtained to estimate the model's parameters, which in this case were:

(a) The probability of each latent class, $\pi_Y(c),\ c = 1,\ldots,C.$

(b) The conditional response probabilities of each manifest variable in each latent class, $\pi_{X_i/Y(c)}(x_i),\ i = 1,\ldots,p;\ c = 1,\ldots,C;\ x_i = 1,\ldots,I_i.$

The posterior analysis of the latent class model examines the data on the individuals of a specific class, using the Y distribution conditioned to X, $\pi_{Y/X(x)}(c) = P(Y = c/X = x)$, called posterior probability distribution. In practice, each individual is placed in the latent class in which this probability is greater. The fundamental assumption is the Local, or Conditional, Independence principle, which establishes that the manifest variables are mutually independent, given the latent variable's fixed value. Basically, this assumption indicates that any association observed between the manifest variables is measured or explained by the latent variables (Magidson & Vermunt, 2004). According to this principle, probability $P(X = x/Y = c)$ is expressed by:

$$\pi_{X/Y(c)}(x) = \prod_{i=1}^{p}\pi_{X_i/Y(c)}(x_i)$$

or equivalent to,

$$\pi_X(x) = \sum_{c=1}^{C}\pi_Y(c)\pi_{X/Y(c)}(x) = \sum_{c=1}^{C}\pi_Y(c)\prod_{i=1}^{p}\pi_{X_i/Y(c)}(x_i)$$

where the parameters are subject to the following restrictions:

$$\sum_{c=1}^{C} \pi_Y(c) = 1$$

$$\sum_{x_i=1}^{I_i} \pi_{X_i/Y(c)}(x_i) = 1,\ i = 1,\ldots,p;\ c = 1,\ldots,C$$

Latent class analysis was performed to obtain differentiated teacher profiles each of which is defined by the different categories of dissimilarity to the G-Rubric software. Finally, the range of profiles obtained were verified by correspondence analysis using SPSS software.

4. Results

Table 1 presents the demographic characteristics of the sample according to degree of dissimilarity. The majority of individuals registered medium dissimilarity (46.7%), with the percentages for high and low dissimilarity the same. The sample distribution for sex was equal between men and women; the majority of individuals were aged 30-50, and the teachers' years of experience was between 0-5 years or 10-20 years.

Table 1. *Demographic Characteristics of the Sample and Dissimilarity*

	Frequency	Percentage
Dissimilarity		
Low	8	26.7%
Moderate	14	46.7%
High	8	26.7%
Sex		
Male	14	46.7%
Female	16	53.3%
Age		
[20,30]	6	20.0%
(30,50]	16	53.3%
More than 50 years	8	26.7%
Experience		
[0,5]	10	33.3%
(5,10]	2	6.7%
(10,20]	10	33.3%
More than 20 years	8	26.7%

Table 2 presents a summary of the estimated models. The L2 statistic indicates the associative quantity between the variables unexplained after model estimation, assuming that a Chi-squared distribution is continued.

Table 2. *Summary of Estimated Models*

		LL	BIC(LL)	Npar	L^2	df	p-value	Class.Err.
Model1	1-Cluster	-120.7966	268.8027	8	76.7044	22	5.60E-08	0
Model2	2-Cluster	-106.7884	257.7923	13	48.688	17	6.70E-05	0.0625
Model3	3-Cluster	-97.4865	256.1945	18	30.0843	12	0.0027	0.034
Model4	4-Cluster	-95.2711	268.7697	23	25.6534	7	0.00058	0.0494

In this case, model 3 presents the lowest BIC (LL) value, thus our model is formed of three clusters. Table 3 shows the model's estimated parameters. For the dissimilarity variables of age and teaching experience, the associated p-value is less than 0.05, thus rejecting the hypothesis that the effects associated with each of these variables is zero. The only p-value more than 0.05 is for the sex variable, which could be significant to 10%. Table 3 presents the R2 value, which indicates which variability percentage of the indicators is explained by the three-cluster model.

Table 3. *Parameter Estimates*

	Cluster1	Cluster2	Cluster3	Wald	p-value	R^2
Dissimilarity	-0.054	-2.7724	2.8264	10.6379	0.0049	0.5729
Sex						
Male	1.979	-0.6364	-1.3426	4.7106	0.0950	0.7295
Female	-1.979	0.6364	1.3426			
Age	-0.6006	-1.8538	2.4544	8.8540	0.0120	0.4356
Experience	0.8528	-2.0436	1.1908	6.3616	0.0420	0.7114

It is observed that 57.29% of the variability in the dissimilarity variable is explained by the model, while the explained variability for sex, age and experience is 72.95%, 43.56% and 71.14%, respectively, which are high values.

Table 4 presents each cluster's profile: cluster 1 is formed of 41.26% of cases, with clusters 2 and 3 containing 34.48% and 24.26%, respectively. The highest category for each variable is highlighted in each cluster. The lowest dissimilarity, the one in which the individuals' scores were closest to the system's, is found in cluster 2. The individuals who form this cluster are women aged 30-50 with between 0-5 years' teaching experience. In contrast, cluster 3 contains those whose scores differed most from the system's, formed of individuals, mainly women, aged 50 or over, with more than 20 years' teaching experience. Cluster 1 consists of individuals whose profile reflected their middling scoring discrepancy against the system.

Table 4. *Profile of Each Cluster*

	Cluster1	Cluster2	Cluster3
Cluster Size	0.4126	0.3448	0.2426
Indicators			
Dissimilarity			
Low	0.0950	**0.6586**	0.0014
Medium	**0.7353**	0.3363	0.1953
High	0.1697	0.0051	**0.8033**
Mean	2.0747	1.3466	2.8019
Sex			
Male	**0.9724**	0.1588	0.0440
Female	0.0276	**0.8412**	**0.9560**
Age			
[20.30]	0.1457	0.4043	0.0018
(30.50]	**0.7109**	**0.5633**	0.1890
>50 years	0.1434	0.0325	**0.8092**
Mean	1.9977	1.6282	2.8073
Experience			
[0.5]	0.0474	**0.8946**	0.0218
(5.10]	0.0718	0.0748	0.0463
(10.20]	**0.5113**	0.0294	0.4628
>20 years	0.3696	0.0012	**0.4691**
Mean	3.2030	1.1372	3.3791

Figure 2 also represents the differences between the three groups. In terms of dissimilarity, there is considerable divergence between the three clusters, as there is for the variables. Cluster 1 differs from clusters 2 and 3 for sex; cluster 3 stands out for age, and cluster 2 for experience.

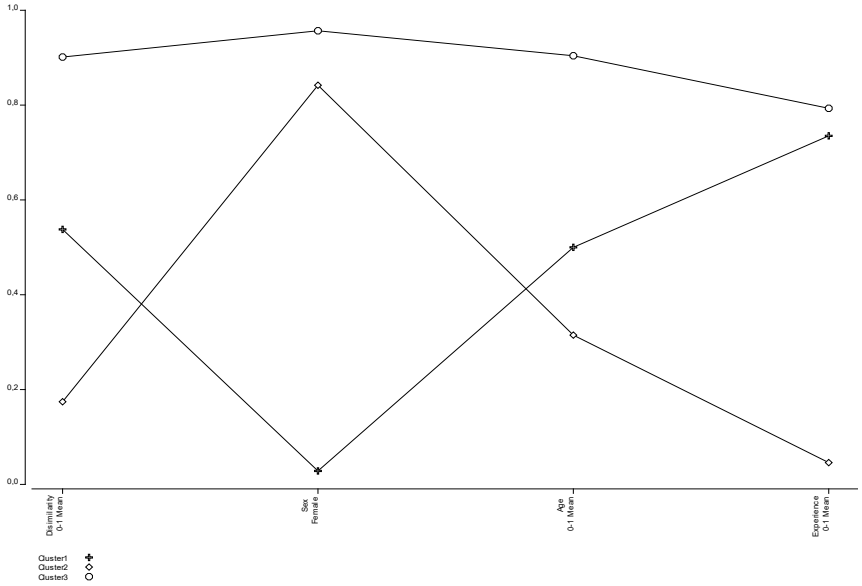

Figure 2. *Graph of Profiles*

Since the aim of the research was to identify a teacher typology that most closely matched the system for assessing style and content in students' written work, and given that the dissimilarity variable could measure the differences between teacher and system scoring for correction, it is cluster 2 that best represents teacher-system similarity. Two conclusions: the differences / similarities between the scores by the individuals and the system can largely be explained by the variables of age and experience; the subjects whose scores most closely match the system's are women aged 30-50 with 0-5 years' teaching experience. To verify the results, a correspondence analysis was performed using SPSS. Table 5 presents a summary of the model with its two dimensions. Inertia indicates the importance of each dimension, the first dimension being more important than the second, as it explains 57.6% of the variability of the variables against 45.7% for the second.

Table 5. *Summary of the Model*

Dimension	Cronbach's Alpha	Explained variance		
		Total (Eigenvalues)	Inertia	% variance
1	.755	2.304	.576	57.598
2	.603	1.826	.456	45.646
Total		4.130	1.032	
Media	.688[a]	2.065	.516	51.622

a. Total Cronbach's Alpha is based on the total eigenvalue.

Figure 3 presents the different categories of the variables. The low category of the dissimilarity variable is associated to the 20-30 age group with 0-5 years' teaching experience and, to a lesser extent, women. Thus, the results obtained are corroborated by the latent class cluster analysis.

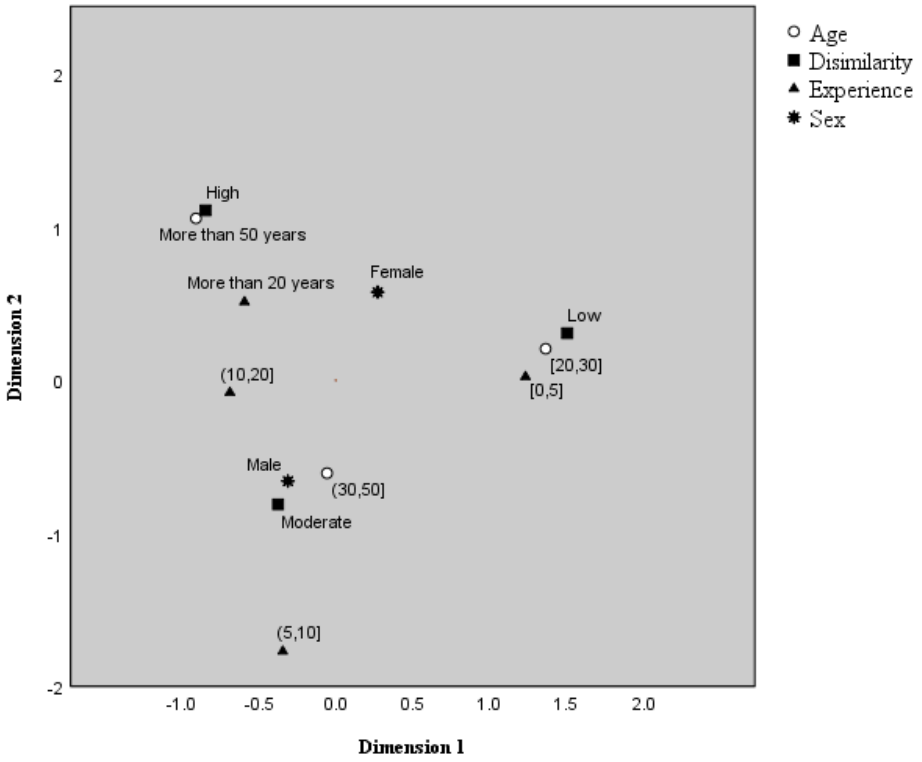

Figure 3. *Representation of Category Points for Each Variable*

5. Discussion

The results of this investigation aim to show which demographic profile of active high school teachers most closely resembles the scoring of the G-Rubric automatic correction tool for assessing students' written production. The results showed a first cluster (24.26%), comprised of teachers whose correction scores exhibited a significant correlation with the tool, represented the quartile of younger and less experienced teachers. This stands in contrast to a second cluster (34.48%), characterized by "high" dissimilarity, which consisted of older and more experienced teachers whose corrections deviated notably from the system, yielding scores lower than those produced by the tool.

This leaves the majority of teachers occupying the middle ground (41.26%), which means that good reliability cannot be confirmed, in terms of the traditional way of measuring this value. The results confirm the findings of other authors (Toranj & Ansari, 2012; Tsai, 2012; Wang & Brown, 2008; Zhang, 2013), who found only marginal correlations between the two correction types.

In this sense, the results of this study confirm that the demographic variables of sex, age and years of teaching experience influence the way students' work is corrected. Women aged 30-50 with between 5-10 years' experience scored most similarly to the correction assessments generated by G-Rubric. Bridgeman et al. (2012), Burstein & Chodorow (1999) and Chodorow & Burstein (2004) had already detected differences between human correctors that related to demographics such as origin and even cultural background. In terms of experience, the results in this study confirm the findings of Royal-Dawson & Baird (2009), who stated that a minimum of three years' experience in the profession was sufficient to reach high levels of reliability, similar to the one obtained with automatic correctors. The results of our study also show that the sex variable relates to human correction scoring that is similar to the software when age and experience intervene, as indicated by cluster 2. High dissimilarity in cluster 3, mainly represented by women, shows that the sex variable in itself has no relation to high levels of concordance with the automatic correction software. This is confirmed by Bridgeman, Trapani and Attali (2012), who found that the sex variable did not exert a significant influence on assessment scores. The opposite is true of age and experience, which mainly explain the similarities to, and the differences from, the instrument.

The results of the corrections made by the participants in this study demonstrate low reliability among teachers, similar to Arnal-Bailera, Muñoz-Escolanoand Oller-Marcén (2016), and Fernández-Alonso et al. (2019) who stated that this was far from unusual. This study's "work experience" and "age" variables that impact on assessments which are similar to G-Rubric's are not the only elements to consider. As mentioned, the dearth of knowledge on teachers' cognitive processes at work during assessment (Zhang, 2013), and the influence of external factors (Brackett, Floman, Ashton-James, Cherkasskiy & Salovey, 2013; Valenti, Neri & Cucchiarelli, 2003; Wohlpart et al., 2008; Zhang, 2013), are variables to consider as impacting on assessments made by human correctors.

Although these data contrast with most of the literature that supports the use of automatic correctors for their reliability (Attali & Burstein, 2006; Bennet, 2010; Bridgeman, 2009; Bridgeman et al., 2012; Crossley et al., 2019; Powers et al., 2000; Shermis & Hamner, 2013; Wang et al., 2008; Warschauer & Ware, 2006; Wohlpart et al., 2008), it is important to note that these investigations apply to software that corrects written work in English, which is structurally less complex than Spanish, thus any Spanish version of these instruments would need to be adapted to deal with such complexities. There are also dissenting voices that doubt the reliability of automatic correctors when compared to human correctors because few studies have measured teachers' cognitive processes when assessing written output, which makes it impossible to validate all the human corrections made in the traditional way (Zhang, 2013).

The literature is unanimous on the use of automatic correctors to assist teachers in correcting students' work (Bridgeman et al., 2012; Chen et al., 2008; Jorge-Botana. Luzón, Gómez-Veiga & Martín Cordero, 2015; Rudner & Gagne 2001; Rudner et al., 2005; Santamaría-Lancho et al., 2018; Toranj & Ansari, 2012; Warschauer & Ware, 2006; Wohlpart et al., 2008; Zhang, 2013). The medium dissimilarity revealed by the Euclidean distances between each teacher and the system represents the largest group of teachers in the study; this middling position could mean that it is these teachers who most benefit from the assistance of G-Rubric when assessing students' work. On the one hand, this could be due to the high concordances (low dissimilarity), which implies less need to use this instrument since correction criteria between both machine and human correctors are similar.

On the other hand, the considerable discrepancies explained by high dissimilarity in experienced teachers imply that assessment criteria differ greatly between teachers and the system. This result has been previously identified in the scientific literature (Bol et al., 1998), where experienced teachers (20 years or more) used alternative methods of assessment more often than least experienced teachers (6 years or less). This could reflect an apparent difficulty in using the instrument as an effective complement to the way corrections are made by this group of teachers, considering how the system currently functions.

The teachers in the medium dissimilarity group, however, are more likely to incorporate the instrument as an efficient way to complement their correction. Powers et al. (2002) explained how teachers were aware of the software's strengths and weaknesses and could confidently delegate to the system certain aspects of correction while they focused on areas such as creativity, originality of thought and structural features that might have gone unnoticed by the instrument (Bridgeman et al., 2012; Pérez-Marín et al., 2006; Rudner & Gagne, 2001).

6. Conclusion

The automatic correction tools available today for evaluating short written answers and summaries in Spanish are not very advanced technologically. There are not enough programs or empirical evidence to ensure a reasonable degree of accuracy across a range of contexts and educational stages. The quality and accuracy of the systems need to improve considerably, one obstacle being that the linguistic characteristics of Spanish are more complex than English, particularly in morphosyntactic alignment. This study has shown that G-Rubric is one of the most promising automatic correction systems for assessing short written answers and summaries, although it is at the experimental stage and its results are still being analyzed. Based on the results obtained in this study, G-Rubric can be supportive in the evaluation of summaries and as a tool to facilitate feedback in learning, but not as a substitute for human evaluation. The degrees of similarity and disparity have varied significantly with respect to teachers' professional experience and age, which should be taken into consideration when applying it across different educational stages.

How to assess open answers and reading comprehension in the form of written summaries are challenges for artificial intelligence and the language sciences in the next few years. Machine correction in education needs to span a wide variety of contexts, as linguistic and communicative competence are basic skills required in pre-university and

university education. Good integration of software in education platforms (PLE, MOOC, LMS, etc.) can help teachers and students by providing the latter with important feedback to enable them to progress in their learning and to verify their understanding of subject matter learned. Students can work on improving a competence, so it is essential for them to receive accurate feedback to detect errors and enhance understanding of texts, and to write better. Teachers can benefit from the potential of learning analytics to get a clearer picture of a student's progress. The study has been contextualized in Spanish educational centers, but due to the high degree of internationalization of the Spanish language, it would be desirable to replicate it in other Spanish-speaking countries to determine its effectiveness in relation to the influence of other socio-educational contexts and teaching profiles. Likewise, it would be desirable to establish comparisons between software programs such as G-Rubric and the results obtained by generative artificial intelligences in the evaluation and feedback processes of open-ended questions and summaries.

Acknowledgments

This work has been developed within the framework of the "Proof of Concept" Project entitled: "GAUBI-ORTO. Evaluation of a digital model and application for sustainable, ubiquitous, and gamified learning of Spanish spelling in Primary Education" (PDC2022-133185-I00). State Program to Promote Scientific-Technical Research and its Transfer, of the State Plan for Scientific and Technical Research and Innovation. Ministry of Science and Innovation (Spain).

Information on Informed Consent or any Data Privacy Statements

Informed consent was obtained from all participants prior to their inclusion in this study. Participants were provided with a written informed consent form, which included information about the purpose of the study, the procedures involved, and any potential risks or benefits associated with participation. Participants were informed that their participation was voluntary and that they could withdraw from the study at any time without penalty. By signing the informed consent form, participants indicated that they had read and understood the information provided and had agreed to participate in the study. The study protocol was conducted in accordance with the principles of the Declaration of Helsinki.

Conflict of Interest

The authors declare no conflicts of interest.

References

ACARA NASOP research team. (2015). An evaluation of automated scoring of NAPLAN persuasive writing. *Acara Australian Curriculum Assessment and Reporting Authority*, *30*. https://nap.edu.au/_resources/20151130_ACARA_research_paper_on_online_automated_scoring.pdf

Amorim, E., & Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 94-102). Association for Computational Linguistics, Valencia, Spain. https://doi.org/10.18653/v1/E17-4010

Arnal-Bailera, A., Muñoz-Escolano, J. M., & Oller-Marcén, A. M. (2016). Characterization of behavior of correctors when grading mathematics tests. *Revista de Educación, 371*, 35-60. doi:10.4438/1988-592X-RE-2015-371-307

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3), 1-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2006). Item selection rules in a computerized adaptive test for the assessment of written English. *Psicothema*, *18*(4), 828-834.

Barrett, C.M. (2015). *Automated essay evaluation and the computational paradigm: Machine scoring enters the classroom* [Unpublished doctoral dissertation]. University of Rhode Island. https://doi.org/10.23860/diss-barrett-catherine-2015

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The analysis and interpretation of multivariate data for social scientists*. Chapman & Hall.

Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 319-341. https://doi.org/10.1080/0969594X.2011.555329

Benítez, M. H., & Lancho, M. S. (2016). G-Rubric: Una aplicación para corrección automática de preguntas abiertas. Primer balance de su utilización. In *Nuevas perspectivas en la investigación docente de la historia económica* (pp. 473-494). Editorial de la Universidad de Cantabria.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2-3), 70-91. https://doi.org/10.1080/15366367.2010.508686

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, *6*(1), 1-47. https://ejournals.bc.edu/index.php/jtla/article/view/1631

Blázquez, M., & Fan, C. (2019). The efficacy of spell check packages specifically designed for second language learners of Spanish. *Pertanika Journal of Social Sciences & Humanities – JSSH, 27*(2), 847-863.

Blumenstein, M., Green, S., Fogelman, S., Nguyen, A., & Muthukkumarasamy, V. (2008). Performance analysis of GAME: A generic automated marking environment. *Computers & Education*, *50*(4), 1203-1216. https://doi.org/10.1016/j.compedu.2006.11.006

Bol, L., Stephenson, P., O'connell, A., & Nunnery, J. (1998). Influence of experience, grade level, and subject area on teachers' assessment practices. *Journal of Educational Research, 91*, 323-330. https://doi.org/10.1080/00220679809597562

Brackett, M. A., Floman, J. L., Ashton-James, C., Cherkasskiy, L., & Salovey, P. (2013). The influence of teacher emotion on grading practices: A preliminary look at the evaluation of student writing. *Teachers and Teaching*, *19*(6), 634-646. https://doi.org/10.1080/13540602.2013.827453

Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann, & J. Bjornsson (Eds.), *The transition to computer-based assessment* (pp. 39-44). Office for Official Publications of the European Communities.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Burstein, J., & Chodorow, M. (1999). *Automated essay scoring for normative English speakers*. Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, Maryland. https://aclanthology.org/W99-0411

Chen, C. F. E., & Cheng, W. Y. E. C. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94-112.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays*. ETS.

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research*, *11*(2), 251-270. https://doi.org/10.17239/jowr-2019.11.02.01

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In *Assessment and teaching of 21st century skills* (pp. 143-230). Springer. https://doi.org/10.1007/978-94-007-2324-5_4

da Cunha, I. (2020). Una herramienta TIC para la redacción del Trabajo de Fin de Grado (TFG). *UA Revistes Cientifiques, 34,* 39-72. https://doi.org/10.14198/ELUA2020.34.2

Díez-Arcón, P., & Martín-Monje (2021). G-Rubric: The use of open technologies to provide personalised feedback in languages for specific purposes. In *EDULEARN21 Proceedings* (pp. 2635-2643). IATED. https://doi.org/10.21125/edulearn.2021.0574

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1), 1-36. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379-393. doi: 10.1177/0146621606288890

Ericsson, P. F., & Haswell, R. H. (2006). *Machine scoring of student essays: Truth and consequences*. All USU Press Publications.

Fernández-Alonso, R., Woitschach, P., & Muñiz Fernández, J. (2019). Rubrics do not neutralize Raters' effects: A many-faceted Rasch model estimation. *Revista de Educación, 386*. 89-112. http://dx.doi.org/10.4438/1988-592X-RE-2019-386-428

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 197-202. https://doi.org/10.3758/BF03204765

Hashemian, M., & Farhang-Ju, M. (2018). Effects of metalinguistic feedback on grammatical accuracy of Iranian field (in)dependent L2 learners' writing Ability. *Journal of Research in Applied Linguistics, 9*(2), 141-161. https://doi.org/10.22055/rals.2018.13797

He, Y., Hui, S. C., & Quan, T. T. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers & Education*, *53*(3), 890-899. https://doi.org/10.1016/j.compedu.2009.05.008

Hoomanfard, M. H., Jafarigohar, M., Jalilifar, A., & Masum, S. M. H. (2018). Comparative study of graduate students' self-perceived needs for written feedback and supervisors' perceptions. *Journal of Research in Applied Linguistics, 9*(2), 24-46. https://doi.org/10.22055/rals.2018.13792

JISC (Joint Information Systems Committee). (2007). *Effective practice with e-assessment: An overview of technologies, policies and practice in further and higher education*. http://www.jisc.ac.uk/media/documents/themes/elearning/effpraceassess.pdf

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.

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing, 18*(1), 62-84. https://doi.org/10.1016/j.asw.2012.10.001.

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in education: Principles, Policy & Practice*, *10*(3), 295-308.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284.

Lazarsfeld, P. F. (1950). *The logical and mathematical foundation of latent structure analysis and the interpretation and mathematical foundation of latent structure analysis*. In S.A. Stouffer et al. (Eds.), *Measurement and prediction* (pp. 362-472), Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis.* Houghton Mill.

Magidson, J., & Vermunt, J.K. (2004). Latent class models. In: Kaplan, D. (ed.) *The Sage handbook of quantitative methodology for the social sciences* (pp. 175-198). Sage Publications, Thousand Oakes.

Martín-Monje, E., & Barcena, E. (2024). Tutor vs. automatic focused feedback and grading of student ESP compositions in an online learning environment. *Journal of Research in Applied Linguistics*, *15*(2), 22-42. https://doi.org/10.22055/rals.2024.45636.3198

Mirzaee, A., & Tazik, K. (2014). Typological description of written formative feedback on student writing in an EFL context. *Journal of Research in Applied Linguistics, 5*(2), 79-94. https://rals.scu.ac.ir/article_11013_1201.html

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, *56*(2), 337-345. https://doi.org/10.1016/j.compedu.2010.07.013

Perelman, L. (2012). Mass-market writing assessments as bullshit. In N. Elliot & L. Perelman, *Writing assessment in the 21st century: Essays in honor of Edward M. White*, (pp. 425-437). Hampton Press.

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hamner: Contrasting state-of-the-art automated scoring of essays: Analysis. *The Journal of Writing Assessment,* *6*(1), 1-10. http://journalofwritingassessment.org/article.php?article=69

Pérez-Marín, D., Alfonseca Cubero, E., & Rodríguez Marín, P. (2006). ¿Pueden los ordenadores evaluar automáticamente preguntas abiertas? *Novátic*, *50*, 50-53.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. *ETS Research Report Series*, *2000*(2), 1-23. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103-134. https://doi.org/10.1016/S0747-5632(01)00052-8

Redecker, C. (2013). *The use of ICT for the assessment of key competences*. Publications Office of the European Union. doi:10.2791/87007

Royal-Dawson, L., & Baird, J. A. (2009). Is teaching experience necessary for reliable scoring of extended English questions?. *Educational Measurement: Issues and Practice*, *28*(2), 2-8. https://doi.org/10.1111/j.1745-3992.2009.00142.x

Rudner, L., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer*. ERIC Digest.

Rudner, L., Garcia, V., & Welch, C. (2005). An evaluation of Intellimetric™ essay scoring system using responses to GMAT AWA prompts. *McLean, VA: GMAC*. https://www.gmac.com/~/media/Files/gmac/Research/research-report-series/RR0508_IntelliMetricAWA.pdf

San Mateo, A. (2016). A bigram corpus used as a grammar checker for Spanish native speakers. *Revista Signos*, *49*(90), 94-118. http://dx.doi.org/10.4067/S0718-09342016000100005

Santamaría-Lancho, M., Hernández, M., Sánchez-Elvira, Á., Luzón, J. M., & Jorge-Botana, G. (2018). Using semantic technologies for formative assessment and scoring in large courses and MOOCs. *Journal of Interactive Media in Education*, *2018*(1), 12. http://doi.org/10.5334/jime.468

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays: Analysis. In M. D. Shermis, & J. Burstein, *Handbook of automated essay evaluation*, (Chapter 19). Routledge. https://doi.org/10.4324/9780203122761

Soleimani, H., & Rahmanian, M. (2014). Self-, peer-, and teacher-assessments in writing improvement: A study of complexity, accuracy, and fluency. *Journal of Research in Applied Linguistics,* *5*(2), 128-148. https://rals.scu.ac.ir/article_11016.html

Toranj, S., & Ansari, D. N. (2012). Automated versus human essay scoring: A comparative study. *Theory and Practice in Language Studies*, *2*(4), 719-725. https://doi.org/10.4304/tpls.2.4.719-725

Tsai, M. H. (2012). The consistency between human raters and an automated essay scoring system in grading high school students' English writing. *Action in Teacher Education*, *34*(4), 328-335. https://doi.org/10.1080/01626620.2012.717033

Usener, C. A., Gruttmann, S., Majchrzak, T. A., & Kuchen, H. (2010). Computer-supported assessment of software verification proofs. In *2010 International Conference on Educational and Information Technology* (Vol. 1, pp. V1-115). IEEE. doi: 10.1109/ICEIT.2010.5607766.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), 319-330. https://doi.org/10.28945/331

Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. *International Journal of Educational Technology in Higher Education, 18*, 33. https://doi.org/10.1186/s41239-021-00269-8

Vázquez-Cano, E., Ramírez-Hurtado, J. M., & Sáez-López, J. M., & López-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity, 49*, 101380. https://doi.org/10.1016/j.tsc.2023.101380

Vermunt, J. K., & Magidson, J. (2002). *Latent class cluster analysis*. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class models* (pp. 89-106). Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2003). *Addendum to Latent Gold user's guide: Upgrade for version* 3. Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent Gold 4.0: Basic and advanced*. Statistical Innovations, Inc.

Villena, J., González, B., González, B., & Muriel, M. (2002). STILUS: Sistema de revisión lingüística de textos en castellano. *Procesamiento del Lenguaje Natural*, *29*, 305-306. https://rua.ua.es/dspace/handle/10045/1759

Wang, H. C., Chang, C. Y., & Li, T. Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, *51*(4), 1450-1466. https://doi.org/10.1016/j.compedu.2008.01.006

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, *8*(4), 310-325.

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, *45*(4), 769-774. https://doi.org/10.5054/tq.2011.272525

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22-36. https://doi.org/10.1080/15544800701771580

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language teaching research*, *10*(2), 157-180. https://doi.org/10.1191/1362168806lr190oa

Wohlpart, A. J., Lindsey, C., & Rademacher, C. (2008). The reliability of computer software to score essays: Innovations in a humanities course. *Computers and Composition*, *25*(2), 203-223. https://doi.org/10.1016/J.COMPCOM.2008.04.001

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2), 1-11. https://www.ets.org/research/policy_research_reports/publications/periodical/2013/jpdd.html