IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388, p- ISSN: 2320-737x Volume 12, Issue 4 Ser. III (Jul. – Aug. 2022), 30-40
www.iosrjournals.org

The effectiveness of Behaviorally Anchored Rating Scales in writing skill evaluation

Antonios Ventouris
School of Italian Language and Literature, Aristotle University of Thessaloniki, Greece

Abstract:
***Background**: The present study proposes a type of Behaviorally Anchored Rating Scale (BARS) for the writing skills assessment evaluation in foreign language education. The commonly used rating scales in language education are numeric. However, numbers can be used only for measurement and counting purposes which are objective procedures, when assessment necessarily involves subjective judgements (Mc Namara, 1996:117). In language assessment though, numbers often express an arbitrary result of language performance, promoting students' classification and disorienting their interest from learning to rating. The BARS is a scale that combines numeric verbal and descriptive evaluation scales that can provide meaningful information facilitating students' improvement and leading the teacher to the proper choices fo`r their support (Schwab et al., 1975, Aiken, 2005).The BARS performance descriptions serve as anchors and permit the comparison with assesses' performance in order to find the correspondence. BARS are considered to reduce construct-irrelevant variance in performance appraisal ratings (Smith & Kendall, 1963)*
***Materials and Methods**: Using this assumption, experimental research was carried out in Aristotle University of Thessaloniki to investigate the effectiveness of BARS over traditional numeric scales in writing skills assessment. The 16 participants to the research were divided into two groups, the experimental group (EG) and the control group (CG). The students were required to complete 10 assignments of written production as part of a B1 CEFR level Italian language course of 50 hours. The students of the first group were assessed with a BARS scale and the second with a traditional numeric by two different raters. At the end of the course both students and raters answered to a short questionnaire about the efficiency of the scale they used. The descriptive analysis of the results was conducted with spss 24 and was calculated the central tendency and the dispersion of the performances in each assignment to find the progress and the performance level. The results of the questionnaire were also analysed with the aforementioned descriptive methods.*
***Results**: Results showed that BARS was highly effective in improving students' overall performance and qualitative attributes of their writing. BARS' effectiveness was also reported by the students and raters who participated in the follow-up field study. In contrast, Numeric Scale did not appear effective in promoting students' improvement.*
***Conclusion:** BARS seem to be effective in helping students to improve their writing skills in a short time while the traditional numeric scales fail to do this.*
***Key Word**: Behaviorally Anchored Rating Scale (BARS), written skills assessment, numeric scales.*

Date of Submission: 07-08-2022 Date of Acceptance: 22-08-2022

I. Introduction

Even if many researchers, teachers, and educational administrators have expressed many doubts about the adequacy, efficiency, validity, and reliability of the numeric-ordinal scales for assessment purposes in education (Allen, 2005; Cizek, 1996; Cox, 2011; Guskey & Bailey, 2001; Kohn, 1999, 2011; Pulfrey et al., 2011), they still remain dominant all over the world, in all the levels of education. In an educational setting, the most substantial problem concerning this type of scale is that they can only communicate ranking information about the student's products and maybe even their "value", in a rather vague way. The assessment though should provide, mainly to the students but also to the other direct or indirect participants of the learning procedure (teachers, principals, coordinators, parents) specific information about the achievements and the progress of the assessees, which will constitute the basis of certain choices and decisions about their learning itinerary (Abma, 2005; Sax, 1974; Stufflebeam, 1974; Ventouris & Loupaki, 2020). A simple number in an ordinal scale is unable to provide such information and sometimes could mislead the learners from their objectives, desires, and needs. In real life though, numbers can only result from two procedures: measurement and counting. Despite the fact that many scholars associate the term of measurement with educational assessment (Bachman, 1990; Davies et al., 1999; Richards et al., 2002; Sax, 1997), it would be difficult to claim that a typical measurement can be conducted in an evaluative procedure. This is due to the fact that a measurement uses an internationally accepted standard unit of measurement and a standard measurement instrument, and its aim is to determine a characteristic of an object (Milanovic, 1998:152).

Nevertheless, educational assessment involves gathering and analysing data, mainly of a qualitative nature, and then interpreting them so that relative decisions can be made (Abma, 2005; Bachman, 1990; Goldstein & Ford, 2002; Ventouris & Loupaki, 2020). These decisions require a comprehensive information and not a ranking, sometimes based on ambiguous data.

A change is not easy. The current practice is most of the times based on tradition and since the educational community is familiar with the use of grades as expression of an evaluative judgment, a change seems to be problematic (Bowers, 2009; Guskey & Bailey, 2001). In addition, a grade represents an evaluation in a synthetic way, even if it is subjective and insufficient for describing a student's achievement and progress. This brief evaluative expression communicates a conclusion, nevertheless the lack of focus on academic criteria concerning student's achievements, progress and knowledge, since many teachers base their judgement on "hodge-podge" criteria like participation, attendance and behavior (Cizek, 2000; Shepard, 2000, 2019). However, this evidence of assessment not only remains vague and, as a result, of limited usefulness, but it also has a negative social and pedagogical effect on students. As Shepard (2019) states "…grading practices elicit comparisons to classmates and imply a permanent lack of ability when learning targets seem out of reach, grading requirements are an obstacle for every teacher hoping to develop a learning-focused classroom culture". This fact creates a climate of competition between students and demotivates the students who receive low grades.

As grades are commonly used, and a complete change in assessment can result in problems, this study proposes a new Behaviorally Anchored Rating Scale (BARS) that combines different types of assessment expressions and enable interested parties to get a clearer information about the quality and characteristics of an assessee's writing production.

In the following chapter are presented the characteristics of the main types of scales, focusing on BARS.

II. The theoretical background of rating scales

The most widespread typology of rating scales classifies them into four main categories:

A. Nominal, which is a categorical scale, permitting the division of certain items into classes on the basis of a given attribute (e.g., Male – Female). The objective of such a scale is not to determine the quality or quantity of a characteristic, but if an individual or an object falls into a particular category.
B. Ordinal, comprising the levelling of an attribute with respect to each other (e.g., 1st, 2nd, 3rd, and so on).
C. Interval, indicating the numbering of different levels in which the intervals between the levels are equal. An example of this type of scale is the Celsius scale, measuring temperature.
D. Ratio, including classes in ordinal relation and with equal intervals between them which have a known distance. The distinguishing characteristic of this type of scale though, is the existence of an absolute zero point, meaning absence of the trait measured.
(Bachman, 1990, 2004; Davies et al., 1999; Sax, 1997; Thorndike & Thorndike-Christ, 2010)

With the exception of the nominal scales, which can show existence or absence of a trait, all the others use numbers to communicate a certain information, like the ranking of an attribute, trait, or characteristic or its measured quantity. However, in education, ratings and rankings may create anxiety (Ali & Anwar, 2021; Wolf & Smith, 1995; Zeidner, 2007), competition among students and conflicts (Barnes et al., 2014; Nelson & Dawson, 2017). Moreover, numbers can only transmit quantitative data, not comprehensive information about a person's communication and language abilities, especially when they result from a rather arbitrary procedure and not from a real measurement. The focus of language education though is mainly on qualitative data, deriving from complex variables, such as accuracy of a text, or cohesion and coherence, where the point is not the number of the element used but their appropriateness and effectiveness in communicative terms.

According to Aiken (1996: 34) there are two basic types of scales: unipolar and bipolar. On a unipolar scale the rater should indicate the extent to which an assessee possesses a certain behavior or trait. For example, the degree of lexical richness of a text can be rated on a scale from 1 to 7, where 1 is the lowest amount of this characteristic and 7 the highest. In a bipolar scale exist two contrary behaviors or traits, designating two extreme categories and the middle represents equal amounts of both of them (e.g., Poor vocabulary – Rich vocabulary).
In a more analytical classification, rating scales can be divided into:
A. Numerical, where the assessor should assign one of several numbers corresponding to particular descriptions of the characteristic to be rated, such as "Task completion" (see figure 1)

Figure 1: Numerical scale

Task completion	1/5	2/5	3/5	4/5	5/5

B. Semantic differential, proposed by Osgood et al. (1957) permit the identification of connotative meaning of certain concepts or attitudes. This type of scale involves an assessment of a subject, using a seven-grade bipolar scale that has two adjectives at each end (Stoklasa et al., 2019; Ventouris, 2006). Usually, for each object of an assessment are used 10 different semantic differential scales (see figure 2).

Figure 2: Semantic differential scale

Vocabulary		
Poor	_ _ _ _ _	Rich
Inappropriate	_ _ _ _ _	Appropriate
Low token/type ratio	_ _ _ _ _	High token/type ratio
Low density	_ _ _ _ _	High density
Inaccurate spelling	_ _ _ _ _	Accurate spelling
…………………………………………………………………………………………		

C. Graphic, which are one of the most popular scales, having at the two end points, and sometimes at intermediate as well, “graphic descriptions of the magnitude of the designed variable corresponding to those points” (Aiken, 1996). These points, named anchors, should be clear and easily comprehensible from assesses and assessors. An example of a graphic scale is:

Figure 3: Grafic scale with descriptors adopted from CEFR (Council of Europe, 2020)

Coherence and cohesion of the text

A1	A2	A2+	B1	B2
Can link words/signs or groups of words/signs with very basic linear connectors (e.g. “and” or “then”).	Can link groups of words/signs with simple connectors (e.g. “and”, “but” and “because”).	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points. Can form longer sentences and link them together using a limited number of cohesive devices, e.g. in a story. Can make simple, logical paragraph breaks in a longer text.	Can use a limited number of cohesive devices to link their utterances into clear, coherent discourse, though there may be some “jumpiness” in a long contribution. Can produce text that is generally well-organised and coherent, using a range of linking expressions and cohesive devices. Can structure longer texts in clear, logical paragraphs.

D. Standard, comprises a set of standards against which the assessees are compared (Aiken, 1996; Ventouris, 2006). Assessments are based on standards established in increasing order, and the assessor must determine which behavior or characteristic (e.g., performance) reflects the best in the assessee. An example of a standard scale is the following:

Figure 4: Standard rating scale with standards adopted from CEFR (Council of Europe, 2018)

According to the student’s written production she/he appears to be…		
Basic user	Independent user	Proficient user

E. Forced-choice, are consisted of two or more descriptive words, phrases, or statements, closely matched in desirability. The assessor has to indicate the one which applies best to the assessee. In case the scale contains more than two choices, the assessor should indicate the most and the least applicable (Aiken, 1996; Brown, 2016; Brown & Maydeu-Olivares, 2017). For an example of a forced-choice scale, adapted for use in language education, see figure 5.

Figure 5: Force-choice scale with statements adopted from KPG (KPG, 2014)

The student writes fully appropriate texts and achieves task communicative purpose.
The student writes texts partly responding to the communicative purpose
Texts exhibited by student are embodying the required characteristics for the level
Student's texts do not achieve communicative purpose or are inappropriate

F. Behaviorally anchored scales (BARS) aim to set in a more transparent and clear way the meaning of their grades, using detailed examples of behaviors for each of them. By using descriptions, this type of scale allows the reader to fully comprehend each point it encompasses, reducing the subjectivity of the grades and the terms it includes (Aiken, 1996), attempting to "capture performance in multidimensional, behavior-specific terms" (Schwab et al., 1975: 222). More precisely, in this kind of assessment, raters must determine whether the behavior they observed corresponds to the description (in fact, BARS were initially called Behavioral Expectation Scales) (Paul et al., 2012).

A major advantage of this scale is that it describes anchor behaviors based on real life, in the case of foreign language education, according to specific target communicative skills. Furthermore, by focusing on specific, concrete, observable behavior as a means of defining the dimensions to be judged and anchoring the evaluative continuum, BARS are considered to reduce construct-irrelevant variance in performance appraisal ratings (Smith & Kendall, 1963). As reported in relative studies, the use of BARS has been hailed as a promising way to improve performance evaluations (Campbell et al., 1970; Dunnette, 1966; Hudson, 2005; Jacobs et al., 1980; Schwab et al., 1975).

A disadvantage of BARS indicated in the relative literature (Borman, 1986; Hudson, 2005) is that the specific behavior descriptions, serving as anchors, may sometimes not match to the ones of the assessees. To address this issue, Borman (1986) introduced the Behavioral Summary Scale (BSS) as an alternative to the BARS scale. In this type, the scales anchor the performance to fewer specific behaviors that reflect a more generalized ability based on various specific incidents at a common level. Another alternative proposed by Borman (1986) is Behavior Observation Scale (BOS), in which the assessor indicates the frequency of certain observable behavioral statements using a five-point scale from almost never to almost always (see figure 6)

Figure 6: Behavior Observation Scale (BOS) - statement adopted from CEFR (Council of Europe, 2001)

Can take messages communicating enquiries, explaining problems						
Almost never	1	2	3	4	5	Almost always

Since in language education the standard descriptors of language performance introduced by official institutions are in most cases sufficiently specific (Byrnes et al., 2012; Council of Europe, 2001, 2018), more generic statements would probably fail to communicate clear information to the assessees and the other interested parties and maybe could generate confusion. Moreover, a frequency indicator seems not applicable in language assessment, as assessments are almost always based on specific samples of the assessee's language ability (e.g., a certain written production).

Based on the BARS approach, this study proposes a functional, comprehensive, and formative scale for assessing written production that may be used identically or as a model for the development of other similar scales.

III. Research: An evaluation of e-BARS' effectiveness in assessing writing skills

The present research was stimulated by the need for clear and comprehensive information from the evaluees about the communicative effectiveness, the quality, and the specific characteristics of their written production. These information will allow them to take the appropriate decision leading to the desired improvement.

This study aimed to develop a multifaceted scale that combines a verbal scale with a numerical scale, as well as provide an accurate description of every value to facilitate students' writing skills improvement.

Research methodology

In order to accomplish the aforementioned aim, a research study of empirical and true experimental design (Cash et al., 2016) was conducted at the Department of Italian Language and Literature. The research subject group consisted of 16 undergraduate students of the Department, participating voluntarily in an extracurricular Italian language course. The aim of the course was the development of writing skills in Italian language at the B1 level. This group was divided equally into two subgroups, the experimental group and control group (see table 1).

Table 1: Research subjects

Research subjects' group							
	Experimental group			Control group			
Age/Sex	Male	Female	Other	Male	Female	Other	Total
20-23	1	2	0	0	3	0	6
24-26	0	5	0	1	4	0	10
Total	1	7	0	1	7	0	16

Table 2: Research subject's distribution by semester

Research subjects' group							
	Experimental group			Control group			
Semester	Male	Female	Other	Male	Female	Other	Total
4th	1	5	0	0	6	0	12
6th	0	1	0	0	1	0	2
8th	0	1	0	1	0	0	2
Total	1	7	0	1	7	0	16

The majority of the study participants were females (14/16), enrolled in the 4th semester (12/16), and aged between 24 and 26 (10/16) (see table 2).
The teachers were both females with more than 20 years of experience teaching Italian to adults, and their ages ranged from 45-50 years (see table 3).

Table 3: Teachers' profile

Teachers' profile			
	Age	**Sex**	**Experience**
Teacher 1	45-55	Female	>20 years
Teacher 2	45-55	Female	>20 years

The research was conducted during the fall semester of 2021-2022 with a study load for the students of 36 hours (26 teaching hours and 10 of study). Each group of students produced 10 written tasks, which were rated by two different rater pairs during the course. The first pair assessed the tasks using an e-BARS form (see subchapter 3.2), created for the needs of the research, and the second an ordinal scale of 10 points. The second group had also the possibility to write comments. In addition to the two groups' teachers, two other experienced Italian language teachers (> 10 years of experience), members of the Greek State Certificate for Foreign Language Proficiency (KPG) raters' board conducted the ratings.
Raters for the experimental group were formed during a 1.30-hour seminar on e-BARS.
The research hypotheses of the research were the following:

A. BARS can help the students to:
- improve their writing performance faster than the traditional numeric scales (NS),
- focus to their disadvantages in writing production, providing them specific information.

B. Bars can help the assessors to:
- rate the written productions of the assessees in a more accurate, reliable, and valid way.

At the end of the course the students and the raters were asked to answer in Greek (their mother language) to an e-questionary created with google forms, containing the following open-ended questions:
A. Do you think the assessment's feedback provided to you/the students helped you/them to improve your/their skills in writing?
B. How would you describe the feedback provided to you/the students? (Answer in not more than 50 words)

IV. Results

Experimental research

The 10 tasks assigned to the students of the experimental (EG) and the control group (CG) were assessed by two assessors, in which one of them was their language teacher.
As shown in table 4, until the third assignment, the two groups' performance on task completion does not differ significantly. The EG students present a lower range of performance ratings (1-2) for the second and third assignment, whereas the CG students have a higher standard deviation (SD). The average scores of EG students gradually improved from the fourth assignment until the final assignment, as indicated by the decrease in relative SD and the gradual increase in mean scores. In contrast, the CG displays a relatively stable mean score and SD for each assignment, indicating that students' performances were not evolving.

Table 4: Task completion results

Descriptive Statistics

Group		N	Minimum	Maximum	Mean	Std. Deviation
Experimental	1st Task Completion	8	1	2	1,50	,535
	2nd Task Completion	8	1	2	1,75	,463
	3rd Task Completion	8	1	2	1,88	,354
	4th Task Completion	8	2	3	2,75	,463
	5th Task Completion	8	2	3	2,75	,463
	6th Task Completion	8	2	3	2,87	,354
	7th Task Completion	8	2	3	2,87	,354
	8th Task Completion	8	2	3	2,87	,354
	9th Task Completion	8	3	3	3,00	,000
	10th Task Completion	8	3	3	3,00	,000
	Valid N (listwise)	8				
Control	1st Task Completion	8	1	2	1,50	,535
	2nd Task Completion	8	1	3	1,50	,756
	3rd Task Completion	8	1	3	1,63	,744
	4th Task Completion	8	1	2	1,63	,518
	5th Task Completion	8	1	2	1,75	,463
	6th Task Completion	8	1	3	1,25	,707
	7th Task Completion	8	1	2	1,50	,535
	8th Task Completion	8	1	2	1,75	,463
	9th Task Completion	8	1	3	1,88	,641
	10th Task Completion	8	1	2	1,63	,518
	Valid N (listwise)	8				

Table 5: Production attributes

Descriptive Statistics

Group		N	Minimum	Maximum	Mean	Std. Deviation
Experimental	1st Production	8	2	6	4,13	1,356
	2nd Production	8	4	7	5,50	1,069
	3rd Production	8	6	9	7,25	,886
	4th Production	8	6	9	7,88	,991
	5th Production	8	7	10	8,13	,991
	6th Production	8	6	9	8,37	1,061
	7th Production	8	7	10	8,75	,886
	8th Production	8	7	10	8,75	,886
	9th Production	8	8	9	8,75	,463
	10th Production	8	8	10	8,75	,707
	Valid N (listwise)	8				
Control	1st Production	8	3	7	4,75	1,282
	2nd Production	8	3	8	4,88	1,808
	3rd Production	8	3	6	4,75	1,165
	4th Production	8	4	6	4,88	,835
	5th Production	8	3	7	4,88	1,356
	6th Production	8	1	8	4,50	2,070
	7th Production	8	2	7	4,88	1,553
	8th Production	8	2	7	4,75	1,581
	9th Production	8	4	6	5,25	,886
	10th Production	8	3	7	4,88	1,246
	Valid N (listwise)	8				

Based on table 5, EG students' performance on tasks' requirements has improved and after the sixth assignment, they reach near the higher end of the scale with a mean score of 8,75. In contrast, the CG students' performance does not demonstrate a significant improvement and their mean score in all assignments does not differ significantly. In light of this fact, we can speculate that the kind of feedback provided to EG students led to efficient decision-making about how to improve their writing. In contrast, it didn't appear that ratings of CG students' assignments helped them improve.

In table 6 is indicated a gradual improvement in EG's text cohesion and coherence. Even if their initial performance (1^{st} – 3^{rd} assignment) was relatively low, with a mean score from 3,88 to 4,75, their progress in the use of cohesion devices and techniques was remarkable and in the two last assignments reached almost the highest level (9 – mean 8,88). Contrary, the progress of the CG students did not appear particularly extensive, starting from a mean score of 2,75 and reaching at the three final assignments a mean of 3,25-3,38. It is rather obvious that the CG assignments' ratings provided them with the opportunity to focus on the relative to criterion characteristics of their texts, but a lack of specific information prevented them from making a more substantial advancement.

Table 6: Cohesion and coherence

Descriptive Statistics

Group		N	Minimum	Maximum	Mean	Std. Deviation
Experimental	1st Cohesion/Coherence	8	3	5	3,88	,641
	2nd Cohesion/Coherence	8	3	5	4,63	,744
	3rd Cohesion/Coherence	8	4	5	4,75	,463
	4th Cohesion/Coherence	8	4	7	5,75	1,165
	5th Cohesion/Coherence	8	4	7	6,63	1,061
	6th Cohesion/Coherence	8	6	8	7,38	,744
	7th Cohesion/Coherence	8	5	9	7,75	1,165
	8th Cohesion/Coherence	8	5	9	7,75	1,165
	9th Cohesion/Coherence	8	8	9	8,88	,354
	10th Cohesion/Coherence	8	8	9	8,88	,354
	Valid N (listwise)	8				
Control	1st Cohesion/Coherence	8	2	4	2,75	,707
	2nd Cohesion/Coherence	8	2	4	2,38	,744
	3rd Cohesion/Coherence	8	2	3	2,63	,518
	4th Cohesion/Coherence	8	2	3	2,62	,518
	5th Cohesion/Coherence	8	2	4	2,63	,744
	6th Cohesion/Coherence	8	2	4	3,00	,535
	7th Cohesion/Coherence	8	2	3	2,87	,354
	8th Cohesion/Coherence	8	3	4	3,25	,463
	9th Cohesion/Coherence	8	3	4	3,38	,518
	10th Cohesion/Coherence	8	3	4	3,25	,463
	Valid N (listwise)	8				

Perhaps the most decisive improvement of EG students' texts appears in the vocabulary quality and quantity. Although they did not perform well on the first assignment (mean score 3,63), they gradually reached the top level of the scale (6). It is noteworthy that after the 4th assignment all the students received very positive assessments (5 or 6), and after the 6th assignment the entire group achieved an assessment higher than 5 (6). Conversely, according to the relative assessments, the CGs' texts vocabulary presented a very limited improvement, while the scores' dispersion appears to be high. This means that there were significant differences between CG students' relative performance, which could indicate that the ratings were unable to assist all students (see table 7).

Table 7

Descriptive Statistics

Group		N	Minimum	Maximum	Mean	Std. Deviation
Experimental	1st Vocabulary	8	3	4	3,63	,518
	2nd Vocabulary	8	3	4	3,87	,354
	3rd Vocabulary	8	4	4	4,00	,000
	4th Vocabulary	8	4	5	4,50	,535
	5th Vocabulary	8	5	6	5,75	,463
	6th Vocabulary	8	5	6	5,62	,518
	7th Vocabulary	8	6	6	6,00	,000
	8th Vocabulary	8	6	6	6,00	,000
	9th Vocabulary	8	6	6	6,00	,000
	10th Vocabulary	8	6	6	6,00	,000
	Valid N (listwise)	8				
Control	1st Vocabulary	8	2	4	3,13	,641
	2nd Vocabulary	8	3	4	3,25	,463
	3rd Vocabulary	8	2	4	2,63	,744
	4th Vocabulary	8	2	5	3,25	,886
	5th Vocabulary	8	3	4	3,13	,354
	6th Vocabulary	8	3	4	3,38	,518
	7th Vocabulary	8	2	5	3,50	,926
	8th Vocabulary	8	3	5	3,50	,756
	9th Vocabulary	8	3	5	3,38	,744
	10th Vocabulary	8	3	4	3,25	,463
	Valid N (listwise)	8				

Table 8

Descriptive Statistics

Group		N	Minimum	Maximum	Mean	Std. Deviation
Experimental	1st Language accuracy	8	2	4	2,88	,641
	2nd Language accuracy	8	4	5	4,25	,463
	3rd Language accuracy	8	4	5	4,88	,354
	4th Language accuracy	8	4	5	4,25	,463
	5th Language accuracy	8	4	5	4,88	,354
	6th Language accuracy	8	4	5	4,75	,463
	7th Language accuracy	8	4	5	4,50	,535
	8th Language accuracy	8	4	5	4,88	,354
	9th Language accuracy	8	5	5	5,00	,000
	10th Language accuracy	8	4	5	4,75	,463
	Valid N (listwise)	8				
Control	1st Language accuracy	8	2	4	2,50	,756
	2nd Language accuracy	8	2	4	2,75	,707
	3rd Language accuracy	8	2	4	3,00	,756
	4th Language accuracy	8	2	4	2,75	,707
	5th Language accuracy	8	2	5	3,00	,926
	6th Language accuracy	8	2	4	2,88	,641
	7th Language accuracy	8	2	4	2,88	,641
	8th Language accuracy	8	3	4	3,13	,354
	9th Language accuracy	8	2	4	2,50	,756
	10th Language accuracy	8	2	4	3,50	,756
	Valid N (listwise)	8				

Language accuracy of EG texts shows a more immediate improvement, which can be attributed to the extensive training that Greek students receive in the language institutes where they learn foreign languages (Bella, 2012; Psaltou-Joycey & Sougari, 2010). Due to the specific and clear feedback about the grammar characteristics of their texts, they were able to make the necessary improvements in short time and reach high levels of language accuracy. The relative performance of the CG students' assignments did not change significantly, since the score range is stable in almost all the assignments (2-4) and the mean score exceeds the 3rd level only in one assignment. In addition, it should be noted that the SD of the scores is high, indicating that despite the feedback given, there are significant differences among students.

Field research

As mentioned in chapter 3.1 at the end of the course students and raters were asked to answer to two open ended questions concerning the feedback usefulness and their overall opinion about BARS.

The answers collected were divided into categories, according to their content and meaning and thus resulted the judgements summarized in tables 10 - 13.

Answering the question "do you think the assessment's feedback provided to you/the students helped you/them to improve their skills in writing?" students and raters confirmed the research questions.

As indicated in table 10 all the respondents of the EG group considered the feedback provided to them through BARS very helpful. One student used a relatively restrained expression to state that he received a lot of help, but it does not seem to convey a negative message.

Table 9: EG students and raters

Role/Answer	Yes, without doubt		A lot		No, I do not	
	N	%	N	%	N	%
Students	3	75	1	25	0	0
Raters	2	100	0	0	0	0

On the other hand, 75% of the CG respondents did not find numerical ratings helpful, and only one claimed it had some benefit for him. Raters expressed a similar opinion with 1 to state that numeric scale was not helpful and the other of limited use.

Table 10: CG students and raters

Role/Answer	Yes, without doubt		To some extent		No, I do not	
	N	%	N	%	N	%
Students	0	0	1	25	3	75
Raters	0	0	1	50	1	50

Students' and raters' answers to the question "How would you describe the feedback provided to you/the students?" are synthesised in tables 12 and 13.

Table 11: EG students and raters

Role/Answer	Effective		Helpful programming		Informative		Complicated		Time consuming*	
	N	%	N	%	N	%	N	%	N	%
Students	4	100	0	0	4	100	1	25	0	0

Raters	2	100	2	100	2	100	0	0	1	50

Table 12: CG students and raters

Role/Answer	**Informative**		**Helpful programming**		**Vague**		**Stressful**		**Misleading***	
	N	%	N	%	N	%	N	%	N	%
Students	1	25	0	0	4	100	4	100	3	75
Raters	1	50	1	50	1	25	2	100	1	50

All the students of the EG indicated the BARS effective and informative, and in their answers, they used the expressions: "they are meaningful", "substantially/really helpful", "informative and "they focus on the content of my text". One of the students though found the BARS complicated and needing a lot of concentration from the student. Raters also described BARS effective and informative, and they added they help to course/lesson programming. However, one of them considered BARS to be time consuming.

The CG's students described their experience with NS as vague and stressful and 3 of them considered it misleading. In only one case, the feedback received was deemed to be informative. Numeric feedback was described as stressful by both raters, and as vague and misleading by one. A positive opinion was expressed by one of them, who noted that the numerical feedback was helpful and informative for programming the lesson.

V. Discussion

According to the research data collected, BARS appears to be more efficient than traditional numeric scales.More precisely, the experimental research conducted suggested that students who received systematic feedback with BARS gradually improved their performance on both the overall and the specific communicative goals. In contrast, the students of the CG who received only numerical feedback did not manage to improve their performance at a significant level. This conclusion could be explained by the informative function of BARS, which allowed the students to understand the points of their text that needed to be changed, without opposing certain linguistic choices and limiting their learning autonomy.In language learning, it is particularly important to develop autonomy since this enables the student to not only learn specific linguistic elements, which are useful in situations that are similar to those of a particular school assignment but also develop metacognitive learning strategies and techniques that facilitate their continued language grow. Considering this, the first two research questions seem to be confirmed since the students of the EG improved their performance in writing in a short time and obviously faster than the students of the CG. Furthermore, due to the improvement they presented in the qualitative traits of their productions (cohesion/coherence, vocabulary, accuracy) we can arrive at the conclusion that EG students managed to focus on their disadvantages in writing production.

According to the evidence collected during the research about the assessor's function, their assessment is more accurate and valid than the CG's. This results indirectly from the EG students' progress and directly from the answers they gave both students and assessors to the relative question. EG's students and assessors evaluated BARS as helpful, effective, and informative, while CG's respondents described numerical ratings mainly stressful and not clear. According to these answers we could claim that also the third research question was confirmed.

The present study showed that BARS is effective and beneficial for improving students' writing skills. The information included in it, however, may appear complicated to some students, perhaps since they must devote more time to deducing it than to numerical ratings. This was also the observation made one EG assessor, describing BARS as time consuming.

The results of this research provide evidence about the effectiveness of the BARS type in foreign language writing production assessment. However, further research is needed to sustain the generalisation of the present results.

The use of a relative BARS form for the assessment of other communicative skills would complete this research and would help teachers to pass more easily from the traditional numerical ratings to a more analytical and informative type of evaluation.

Finally, more research is needed to reach a conclusion about BARS reliability and validity, maybe using relative statistical methods.

VI. Conclusion

Considering the data collected from the present research, BARS results efficient for supporting students' writing skills improvement in a constant way and in a short time, while numeric scales seem to fail to achieve a formative objective.

References

[1] A. Kohn, "The Case against Grades," *Educ. Leadersh.*, vol. 69, no. 3, pp. 28–33, 2011, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ963095&site=ehost-live%5Cnhttp://www.ascd.org/publications/educational-leadership/nov11/vol69/num03/abstract.aspx.

[2] A. Kohn, "From degrading to de-grading," *High Sch. Mag.*, vol. 6, no. 5, pp. 38–43, 1999.

[3] G. J. Cizek, "Grades: The final frontier in assessment reform," *NASSP Bull.*, vol. 80, no. 584, pp. 103–110, 1996, doi: 10.1177/019263659608058416.

[4] K. B. Cox, "Putting Classroom Grading on the Table : A Reform in Progress," *Am. Second. Educ.*, vol. 40, no. 1, pp. 67–87, 2011.

[5] C. Pulfrey, C. Buchs, and F. Butera, "Why grades engender performance-avoidance goals: The mediating role of autonomous motivation," *J. Educ. Psychol.*, vol. 103, no. 3, pp. 683–700, 2011, doi: 10.1037/a0023911.

[6] J. D. Allen, "Grades as Valid Measures of Academic Achievement of Classroom Learning, The Clearing House: A Journal of Educational Strategies," *Issues and Ideas*, vol. 78, no. 5, pp. 218–223, 2005, doi: 10.3200/TCHS.78.5.218-223.

[7] T. R. Guskey and J. M. Bailey, *Developing Grading and Reporting Systems for Student Learning*. Thousand Oaks, California: Corwin Press. Inc, 2001.

[8] A. Ventouris and E. Loupaki, "Final Assessment and Testing for Community Interpreter Trainers: A Theore cal Approach," in *Teacher Education for Community Interpreting and Intercultural Mediation: Selected Chapters*, Ljubljana: Ljubljana University Press, 2020, pp. 228–256.

[9] D. L. Stufflebeam, "Alternative Approaches to Educational Evaluation: A Self-Study Guide for Educators," in *Evaluation in Education*, W. J. Popham, Ed. California: McCutchan Publishing Corp., 1974, pp. 116–213.

[10] G. Sax, "The Use of Standardized Tests in Evaluation.," in *Evaluation in Education*, W. J. Popham, Ed. California: McCutchan Publishing Corp., 1974, pp. 243–308.

[11] T. A. Abma, "Responsive evaluation: Its meaning and special contribution to health promotion," *Eval. Program Plann.*, vol. 28, no. 3, pp. 279–289, Aug. 2005, doi: 10.1016/J.EVALPROGPLAN.2005.04.003.

[12] L. F. Bachman, *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press, 1990.

[13] J. C. Richards, J. Platt, and H. Platt, *Dictionary of language teaching & applied linguistics*. Harlow: Longman, 2002.

[14] G. Sax, *Principles of educational and psychological measurement and evaluation*. Wadsworth, 1997.

[15] A. Davies, A. Brown, C. Elder, K. Hill, T. Lumley, and T. McNamara, *Dictionary of language testing*. Cambridge: Press Syndicate of the University of Cambridge, 1999.

[16] M. Milanovic, *Studies in Language Testing (SiLT)*. Cambridge: Cambridge University Press, 1998.

[17] I. L. Goldstein and J. K. Ford, *Training in organizations: Needs assessment, development, and evaluation*. Belmont, CA: Wadsworth/Thomson Learning, 2002.

[18] A. J. Bowers, "Reconsidering grades as data for decision making: more than just academic knowledge," *J. Educ. Adm.*, vol. 47, no. 5, pp. 609–629, 2009, doi: 10.1108/09578230910981080.

[19] G. J. Cizek, "Pockets of resistance in the assessment revolution," *Educ. Meas. Issues Pract.*, vol. 19, no. 2, pp. 16–23, 2000, doi: 10.1111/j.1745-3992.2000.tb00026.x.

[20] L. A. Shepard, "The Role of Assessment in a Learning Culture," *Educ. Resea*, vol. 29, no. 7, pp. 4–14, 2000.

[21] L. A. Shepard, "Classroom Assessment to Support Teaching and Learning," *ANNALS*, p. 683, 2019, doi: 10.1177/0002716219843818.

[22] L. F. Bachman, *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press, 2004.

[23] R. Thorndike and T. Thorndike-Christ, *Measurement and Evaluation in Psychology and Education - Robert M. Thorndike, Tracy M. Thorndike-Christ - Google Libri*, 3rd ed. Boston: Pearson, 2010.

[24] M. Zeidner, "Test Anxiety in Educational Contexts. Concepts, Findings, and Future Directions.," in *Emotion in Education*, P. A. Schutz and R. Pekrun, Eds. Academic Press, 2007, pp. 165–184.

[25] B. J. Ali and G. Anwar, "Anxiety and Foreign Language Learning: Analysis of students' anxiety towards Foreign language learning," *Int. J. English Lit. Soc. Sci.*, vol. 6, no. 3, pp. 234–244, 2021, doi: 10.22161/ijels.63.32.

[26] L. F. Wolf and J. K. Smith, "The Consequence of Consequence: Motivation, Anxiety, and Test Performance," *Appl. Meas. Educ.*, vol. 8, no. 3, pp. 227–242, 1995, doi: 10.1207/s15324818ame0803_3.

[27] R. Nelson and P. Dawson, "Competition, education and assessment: connecting history with recent scholarship," *Assess. Eval. High. Educ.*, vol. 42, no. 2, pp. 304–315, 2017, doi: 10.1080/02602938.2015.1105932.

[28] N. Barnes, H. Fives, and C. Dacey, "Teachers' beliefs about assessment," in *International Handbook of Research on Teachers' Beliefs*, H. Fives and M. Gregoire Gill, Eds. New York, London: Routledge, 2014, pp. 296–312.

[29] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Chicago: University of Illinois Press, 1957.

[30] J. Stoklasa, T. Talášek, and J. Stoklasová, "Semantic differential for the twenty-first century: scale relevance and uncertainty entering the semantic space," *Qual Quant*, vol. 53, pp. 435–448, 2019, doi: 10.1007/s11135-018-0762-1.

[31] A. Ventouris, "University teacher's evaluation by students: A data collection form proposal," Aristotle University of Thessaloniki, 2006.

[32] L. R. Aiken, *Rating scales and checklists : evaluating behavior, personality, and attitudes*. Washington: John Wiley & Sons, 1996.

[33] Council of Europe, *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, vol. 13. Strasbourg: Council of Europe Publishing, 2020.

[34] Council of Europe, *COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT COMPANION VOLUME WITH NEW DESCRIPTORS*. Strasbourg: Council of Europe, 2018.

[35] A. Brown, *Item Response Models for Forced-Choice Questionnaires: A Common Framework*, vol. 81, no. 1. 2016.

[36] A. Brown and A. Maydeu-Olivares, *Modelling forced-choice response formats*, vol. 2–2, no. January. 2017.

[37] KPG, "Script Rater Guide," Athens, 2014. [Online]. Available: https://rcel2.enl.uoa.gr/kpg/files/Script_Raters_Guide_May_2014.pdf.

[38] D. P. Schwab, H. G. Heneman, and T. A. DeCotiis, "Behaviorally Anchored Rating Scales: a Review of the Literature," *Pers. Psychol.*, vol. 28, no. 4, pp. 549–562, 1975, doi: 10.1111/j.1744-6570.1975.tb01392.x.

[39] M. Paul, M. I. Graef, and K. Saathoff, "Developing Behavior-Based Rating Scales for Performance Assessments," no. May, pp. 1–18, 2012.

[40] P. C. Smith and L. M. Kendall, "Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales," *J. Appl. Psychol.*, vol. 47, no. 2, pp. 149–155, 1963, doi: 10.1037/h0047060.

[41] M. D. Dunnette, *Personnel selection and placement.* Wadsworth Publishing Company, 1966.

[42] J. P. Campbell, M. D. Dunnette, E. E. Lawler, and K. E. Weick;, *Managerial behavior, performance and effectiveness*. New York: McGraw-Hill Book Company, 1970.

[43] R. Jacobs, D. Kafry, and S. Zedeck, "Expectations of Behaviorally Anchored Rating Scales," *Pers. Psychol.*, vol. 33, no. 3, pp. 595–640, 1980, doi: 10.1111/j.1744-6570.1980.tb00486.x.

[44] T. Hudson, "Trends in assessment scales and criterion-referenced language assessment," *Annu. Rev. Appl. Linguist.*, vol. 25, pp. 205–227, 2005, doi: 10.1017/S0267190505000115.

[45] W. C. Borman, "Behavior-based rating scales," in *Performance assessment: Methods & applications*, R. A. Berk, Ed. London: Johns Hopkins University Press, 1986, pp. 100–120.

[46] Council of Europe, "The Common European Framework of Reference for Languages : Learning, Teaching, Assessment," *Counc. Eur.*, 2001, doi: 10.1017/S0267190514000221.

[47] H. Byrnes, J. Child, N. Patrizio, and P. Lowe, "ACTFL proficiency guidelines," 2012. [Online]. Available: http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf.

[48] P. Cash, T. Stanković, and M. Štorga, *Experimental design research: Approaches, perspectives, applications*. Springer International Publishing, 2016.

[49] S. Bella, "Pragmatic development in a foreign language : A study of Greek FL requests," *J. Pragmat.*, vol. 44, no. 13, pp. 1917–1947, 2012, doi: 10.1016/j.pragma.2012.08.014.

[50] A. Psaltou-Joycey and A. Sougari, "Greek young learners ' perceptions about foreign language learning and teaching," in *Advances in Research on Language Learning and Teaching: Selected Papers.*, Thessaloniki: The Greek Applied Linguistics Association, 2010, pp. 387–401.

Antonios Ventouris. "The effectiveness of Behaviorally Anchored Rating Scales in writing skill evaluation." *IOSR Journal of Research & Method in Education (IOSR-JRME)*, 12(04), (2022): pp. 30-40.