Measuring Teaching Effectiveness with Behavioral Scales: A Systematic Literature Review

Luis Matosas-López,[1] Rey Juan Carlos University, Spain

Abstract: Behaviorally Anchored Rating Scales or BARS have been used to measure teaching effectiveness in different stages, programs, and instruction modalities. The author's purpose in this article is twofold: to examine, to date, the use of BARS in the educational setting for the assessment of teacher performance and to demonstrate the possibilities of using this type of instruments in the near future. To do so, the author examines the publications on this issue by conducting a systematic literature review in two distinct phases. In the first phase, queries of different key terms are made in the Web of Science (WOS) database. This is followed by a second phase in which the title, abstract, and keywords of each publication are manually examined. The findings reveal an increase in the use of BARS in recent years providing evidence on the high potential for application of this type of instrument for the appraisal of teaching effectiveness. The author concludes that research on the use of BARS for the evaluation of teacher performance is not only an emerging line of study, but also a topic that invites the academic community to reflect on which mechanisms are most appropriate when dealing with a serious task, in the context of educational organization, such as the appraisal of teaching effectiveness. The originality of this work lies in being the first systematic literature review on the use of BARS for the assessment of teacher performance in the educational setting.

Keywords: Teaching Effectiveness, Teacher Performance, Assessment of Teaching, Behavioral Scales, BARS

Introduction

Since the development of the Behaviorally Anchored Rating Scales, also known as BARS, to assess the performance of healthcare professionals (Smith and Kendall 1963), these scales have been used to measure personnel performance in different organizational settings. For example, Landy and Guion (1970) analyzed the motivation and productivity of engineers, while Fogli, Hulin, and Blood (1971) developed BARS to study the performance of salespeople in food retail supermarkets. Williams and Seiler (1973) explore the effort of engineering professionals in the industrial setting, and Arvey and Hoyle (1974) apply these scales to evaluate the efficiency of system analyst-programmers. Bearden and Wagner (1988) use BARS to measure the work of mechanical operators in the navy, while Woods, Sciarini, and Breiter (1998) postulate the use of behavioral scales to evaluate professionals in the tourism sector. Catano (2007) uses BARS to collect decision-making information on the promotion of police officers. MacDonald and Sulsky (2009) use this system to compare the efficiency of management models in Eastern and Western business cultures.

Nevertheless, the measurement of job efficiency is always a difficult task in any organizational setting. When we talk about measuring job efficiency, this issue is generally approached by using two terms: "Evaluation" and "Assessment." Although these terms present certain differential nuances—the evaluation focuses more on performance control, while when speaking of assessment, the focus is not so much on the control but on the performance improvement from a constructive point of view—both are often used simultaneously to refer to the idea of measurement in its broadest sense (Barnokhon 2022; Leguey Galán, Leguey Galán, and Matosas López 2018; Matosas-López and García-Sánchez 2019). This is the sense in which both terms are used in the present study.

[1] Corresponding Author: Luis Matosas-López, Department of Financial Economics and Accounting, Rey Juan Carlos University, Madrid, Spain. email: luis.matosas@urjc.es

The International Journal of Educational Organization and Leadership
Volume 30, Issue 1, 2023, https://thelearner.com

ISSN: 2329-1656 (Print), ISSN: 2329-1591 (Online)
https://doi.org/10.18848/2329-1656/CGP/v30i01/43-58 (Article)

Such systems that are used to measure teaching effectiveness can often be ambiguous and subjective. Likewise, often these processes do not provide explicit descriptions of job dimensions and expected performance levels, and consequently, the measurements are conditioned by the evaluators' interpretation of these parameters. Obviously, this situation generates different scores, even when the ratees present identical behaviors.

The concept of measurement itself can be ambiguous and subjective. So, even the definitions of "teacher effectiveness" and "teacher performance" can present nuances or differences between them. Here, different approaches can be identified; however, according to different reports and authors, while "teacher effectiveness" focuses normally on the efficacy and productivity of teachers in their duties, "teacher performance," typically, focuses on their action from a general point of view (Matosas-López, Leguey-Galán, and Doncel-Pedrera 2019; Medley 1977). In this sense, it is worth mentioning the studies and theories on teacher efficacy or educational effectiveness by authors such as Hoy and Woolfolk (1993), Lynott and Woolfolk (1994), Kitsantas (2012), as well as Reynolds and Teddlie (2000), among others.

BARS appeared in the early 1960s with the purpose of improving the objectivity of ratings in the assessment of job efficiency and mitigating the impact of evaluator interpretations (Matosas-López, Leguey-Galán, and Doncel-Pedrera 2019). To do so, the instrument is defined in behavioral terms, offering concrete examples of actions that illustrate the different levels of performance considered in that work (Bernardin and Smith 1981; Smith and Kendall 1963). Accordingly, the value of these scales, in comparison with other measurement systems, lies in their use of behavioral examples to represent each of the anchor points constituting the scale for each of the job dimensions evaluated. The use of behavioral examples in scale anchor points ensures a more standardized and uniform understanding of the performance level in each job dimension, thereby enabling more consistent, precise, and objective interpretations (Bernardin and Beatty 1984).

Although the first BARS proposal was by Smith and Kendall (1963), the antecedent of this type of measurement is the critical incident technique presented by Flanagan (1954). This technique proposes the structured compilation of behavioral examples that are characteristic of the job evaluated through group interviews and surveys with individuals with experience in the activity analyzed. The closeness between the critical incident technique and BARS is such that Campbell, Dunnette, and Arvey (1973) define BARS as scales based on critical incidents.

Even though the original methodology by Smith and Kendall (1963) has undergone slight variations through the years, the fundamental steps remain as follows. First, a panel of experts in the activity under evaluation describes the job dimensions in detail. Second, a group of individuals directly connected to the activity provides efficient and inefficient critical incidents or behavioral examples for each job dimension. Third, critical incidents are filtered to eliminate duplicate or ambiguous examples. Fourth, another group of individuals relocates the critical incidents to the job dimension for which they were originally formulated, eliminating those that are not correctly reassigned by most participants. Fifth, the critical incidents that survive the previous relocation are rated again by the participants on an ordinal scale. Finally, the researchers select the critical incidents that will serve as anchor points for each performance level in each dimension of the final scale.

Advantages and Drawbacks of BARS

Advantages

Behavioral scales have attracted the interest of researchers and practitioners in many different organizational contexts. Part of the success of the BARS lies in its psychometric advantages over other measurement systems, especially compared with Likert-type instruments, which, while reasonably effective (Vanacore and Pellegrino 2019; Zhao and Gallant 2012), have also raised serious doubts about their suitability for the evaluation of job efficiency (Hornstein 2017; Spooren, Brockx, and Mortelmans 2013).

BARS can reduce both the halo effect and the leniency error. The halo effect is defined as the evaluator's tendency to extrapolate the rating on a particular question to all the survey items (Bernardin 1977). The leniency error, in contrast, is the evaluator's inclination to rate the professional too high or too low on all the questionnaire items (Sharon and Bartlett 1969). Here, several investigations demonstrate that BARS tends to produce a smaller halo effect and leniency error than other types of scales (Bernardin, Alvares, and Cranny 1976; Borman and Dunnette 1975; Campbell, Dunnette, and Arvey 1973). Smith and Kendall (1963) affirm that the filtering and relocation of critical incidents, based on participant agreement, during the construction process is the key to achieving these reductions.

Other benefits of BARS are improvements in validity and reductions in the influence of bias during the assessment (Debnath, Lee, and Tandon 2015; Martin-Raugh et al. 2016; Ohland, Loughry, and Woehr 2012). Murphy and Pardaffy (1989) note that this advantage results from the total disconnection between the scales, which isolates them from biases originating in other dimensions during the construction process.

Along the same lines, many studies suggest that behavioral scales provide indicators of better interrater reliability than those found in other questionnaires (Bearden and Wagner 1988; Debnath, Lee, and Tandon 2015; Williams and Seiler 1973). Bernardin (1977) defines interrater reliability as the degree of agreement reached among raters on various dimensions when evaluating a particular individual. According to Bernardin and Smith (1981), the use of behavioral examples to represent each anchor point of the scale, in addition to the use of clearly independent dimensions, contributes to this improvement.

The benefits of scales with behavioral episodes are such that even some authors state that BARS are technically and psychometrically better than any other measurement instrument (Borman and Dunnette 1975; Goodale and Burke 1975; Matosas-López and Bernal-Bravo 2020). This superiority, in psychometric terms, is often attributed to the rigor of scale construction (Borman 1991), the direct involvement of individuals connected with the activity under evaluation in the instrument design (Bernardin and Beatty 1984), and even the benefits of using terminology familiar to the rater in the final questionnaire (Jacobs, Kafry, and Zedeck 1980).

Drawbacks

Even though behavioral scales have been demonstrated to provide significant benefits in assessing job efficiency, these instruments also present several disadvantages. The literature on BARS points out three problems in this regard: (a) the high investment of time and effort required for the design and construction of the questionnaire; (b) the difficulty in obtaining examples of behavior that is representative of scale midpoints; and (c) the loss of behavioral information suffered throughout the construction process.

Although part of the potential of BARS lies in its meticulous design, this is also one of the drawbacks associated with its use. The considerable time and effort required to gather and filter the critical incidents throughout the successive steps of scale construction can pose a barrier for those interested in these instruments (Goodale and Burke 1975; Stoskopf et al. 2016).

Another problem is the difficulty in reaching substantial degrees of agreement on the critical incidents that illustrate the intermediate anchor points on the scale. Although the levels of agreement on the extreme points—those of lower and higher efficiency—usually present a large quorum, the management of intermediate values represents an important challenge for the researcher (Debnath, Lee, and Tandon 2015; Hauenstein, Brown, and Sinclair 2010).

Finally, the loss of behavioral information during instrument design is also a frequent concern. Matosas-López, Leguey-Galán, and Leguey-Galán (2019) note the constant and substantial loss of information in the construction of these scales. This loss of information is a consequence of the elimination of behavioral examples throughout the screenings carried out during the construction process. This problem has been confirmed in studies reflecting a loss of

up to 80 percent of the critical incidents originally considered for scale construction (Carretta and Walters 1991; Kell et al. 2017; Klieger et al. 2018). This loss of behavioral information, which is the cornerstone of the instrument, also undermines the potential of the scales themselves.

The Use of BARS in the Educational Context

The assessment of job efficiency has become an essential concept in every organizational setting. Therefore, these measurement systems are used in several relevant areas of society: production processes, agriculture sector, food industry, government plans, healthcare programs, urban development, transportation, and so on. There is no organizational context or activity that remains outside these control mechanisms, and of course, education is no exception. In the field of education, the measurement of job efficiency is closely connected to the assessment of quality (Langton 2013; Matosas 2018; Mula-Falcón 2021; Wea and Werang 2020).

In this setting, the assessment of quality can have multiple connotations and approaches (Hernandez-Ramos and Martínez-Abad 2021; Matosas-López et al. 2021). However, two approaches stand out: one focusing on the idea of service (Veciana-Vergés and Capelleras-i-Segura 2004) and the other on student satisfaction (Alvarado-Lagunas, Ramírez, and Téllez 2016). The first considers issues such as teacher competences, the attention of service staff, or facilities, while the second focuses on aspects such as the instructional methods used by the teacher, the level of specialization acquired, or complementary training.

Nevertheless, regardless of the approach, the idea of quality in the educational setting is closely connected with the efficiency of the teaching staff, sometimes even overlapping with the concept of quality in its broadest sense (López-Cámara, González-López, and De Leon-Huertas 2014; Gómez-Gómez and García-Aretio 2016).

In this mission of measuring teaching effectiveness, BARS instruments have been well received and are used in different stages, programs, and modalities of instruction. Bernardin (1977) uses scales with behavioral examples to measure the teaching effectiveness of university professors. Kavanagh and Duffy (1978) use this type of instrument to evaluate a distance education program aimed at improving reading skills. Hom et al. (1982) use this system to appraise teachers of a summer school program. Fernández-Millán and Fernández-Navas (2013) use these scales to evaluate the efficiency of social educators in child protection centers. Martin-Raugh et al. (2016) apply BARS to assess the activity of English and mathematics teachers in a primary school.

These are just some examples of studies that can be found in the literature on the use of BARS in the educational setting to measure the efficiency of teaching staff; however, these initial samples merely illustrate the potential of these scales.

Objective

The objective of the present study is twofold: to examine, to date, the use of BARS in the educational setting for the assessment of teacher performance and to demonstrate the possibilities of using this type of instruments in the near future. To do so, the author examines the publications made on this issue by conducting a systematic literature review in two distinct phases.

In the first phase, queries of different key terms are made in the Web of Science database (hereinafter WOS). During the second phase, the title, abstract, and keywords of each publication are manually examined, to verify its fit to the objective of study. The work concludes with a discussion of the findings obtained and the main conclusions drawn.

The originality of this work lies in being the first systematic literature review on the use of BARS for the assessment of teaching effectiveness, revealing the increase in the use of this type of instruments and also providing evidence on the high potential for application of this type of questionnaires in the context of educational organization.

Methodology

The researcher carried out this systematic literature review in WOS using the so-called Main Collection of the database. This collection uses as a source, among others, the following directories: Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (A&HCI), and Emerging Sources Citation Index (ESCI).

To ensure that the analysis provides an international perspective, the author retrieved the information from the works published in English. Likewise, to ensure that only those publications of greater consistency and depth are included, only journal articles, review articles and book chapters were taken, leaving out of the queries conference proceedings, meeting abstracts, and notes. As for the time horizon concerns, no time limits were established in the queries, which allowed recovering all works on this topic, regardless of its year of publication.

All searches were executed using the query field "Topic." This allowed the researcher to examine the information contained in the title, abstract, and keywords subfields of each work.

In the first place, the author retrieved the publications dealing with the issue of BARS in a broader sense. Following this, the researcher extracted, specifically, those papers on the use of BARS, in the educational setting, for the appraisal of teaching effectiveness.

In the line of previous literature reviews (Ruiz-Ariza and Cruz-González 2021), the author, after the extraction of publications from the database, carried out a manual and comprehensive exploration of the retrieved papers. In this manual review, the researcher revised the title, abstract, and keywords of each publication to verify that the extracted papers met the objective of the study previously defined.

Finally, after completing the phases of automated and manual skimming, the author observed in detail the publications that, in accordance with the number of citations obtained, had had the greatest impact. At this point, the five articles with the highest number of citations were examined.

Results

In the first query, the one related to the publications that address the topic of BARS from a general perspective, terms such us Behaviorally Anchored Rating Scales, Behavioral Scales or Behavioral Episodes—in the US English format—and Behaviourally Anchored Rating Scales, Behavioural Scales or Behavioural Episodes—in the UK English format—were used.

The term BARS was excluded because it overlaps with other terms that had nothing to do with the use of behavioral scales, thus avoiding search results unrelated to the issue under observation from being thrown up. The syntax used by the author in this first query, combining search terms and Boolean OR operator, was the following:

> TS = ("behaviorally anchored rating scales" OR "behaviourally anchored rating scales" OR "behavioral anchored rating scales" OR "behavioural anchored rating scales" OR "behavioral episodes" OR "behavioural episodes" OR " behavioral scales" OR "behavioural scales")

The query carried out with the above syntax yielded a total of 379 publications between the years 1963 and 2022. Since Smith and Kendall (1963) published the first paper on the design and use of behavioral scales, titled "Retranslation of Expectations: an approach to the construction of unambiguous anchors for rating scales," in the *Journal of Applied Psychology*, publications on this topic have experienced an evident growth (Figure 1).

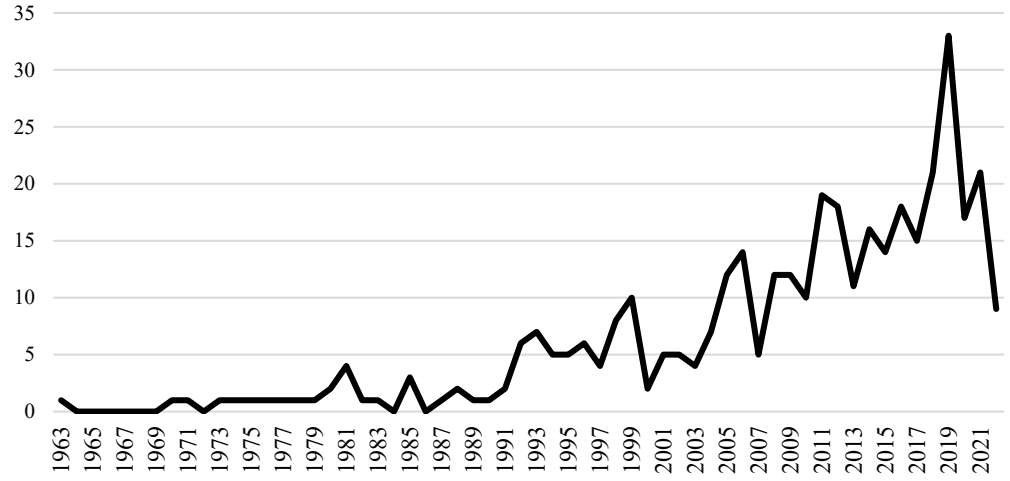

Figure 1: Evolution in the Number of Publications on the Use of BARS in Different Contexts

These papers received a total of 14,072 citations, with an average number 37.63 citations per paper. In the 379 publications retrieved, an h-index of 53 is also observed, indicating that at least 53 of these works received 53 citations or more.

Regarding the thematic categories, the papers on this issue were published mainly in journals on neurosciences, clinical neurology, psychiatry, pediatrics, psychology applied, rehabilitation, psychology clinical, and management. The concentration of publications in categories related to health sciences disciplines reveals that BARS, aside from their potential for measuring job efficiency, have been widely used for the evaluation of disorders and sicknesses in the sanitary field. The entire distribution of publications per thematic category of the journal is presented in Figure 2.

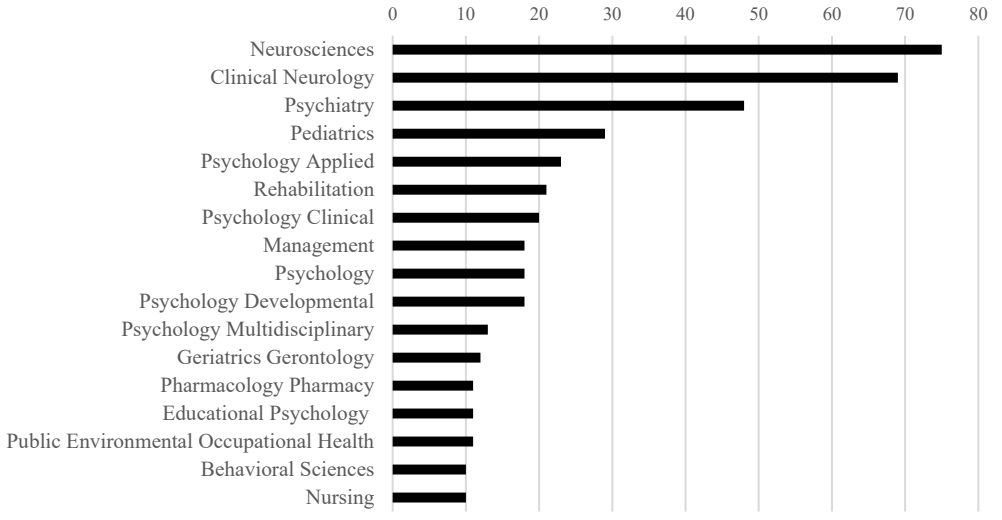

Figure 2: Publications on the Use of BARS, in Different Contexts, per Journal Thematic Category

The Use of BARS in the Educational Context

Once this first general exploration was completed, the author retrieved publications on the use of BARS for the assessment of teaching effectiveness. In the search syntax, the structure used in the first query was taken as a basis to identify the papers focused, in general, on the topic of BARS. This syntax was complemented with two additional extensions. The first to limit the application of BARS in the educational setting, and the second to specify that in this area the scales were used to assess teaching effectiveness. Therefore, the final syntax consists of three parts: the one related to the use of BARS, the one related to their application in educational contexts, and the one related to appraisal of teaching effectiveness. These three parts connected using the Boolean AND operator result in the next syntax:

> TS = ((“behaviorally anchored rating scales” OR “behaviourally anchored rating scales” OR “behavioral anchored rating scales” OR “behavioural anchored rating scales” OR “behavioral episodes” OR “behavioural episodes” OR “ behavioral scales” OR “behavioural scales”) AND (“teacher” OR “professor” OR “school” OR “high school” OR “higher education” OR “university”) AND (“teaching effectiveness” OR “teaching efficiency” OR “teacher performance”))

This query shows the existence of 51 publications that specifically deal with the topic under analyses, dating the first one of them from 1976 (Bernardin, Alvares, and Cranny 1976). After the automated skimming was carried out using the query in WOS, the author proceeded with the manual skimming reviewing titles, abstract, and keywords of each paper. In this phase, from the 51 works identified initially, only 44 fitted the topic of study accurately. Consequently, only 11.61 percent of the publications retrieved in the first query focused specifically on the application of BARS for the assessment of teaching effectiveness.

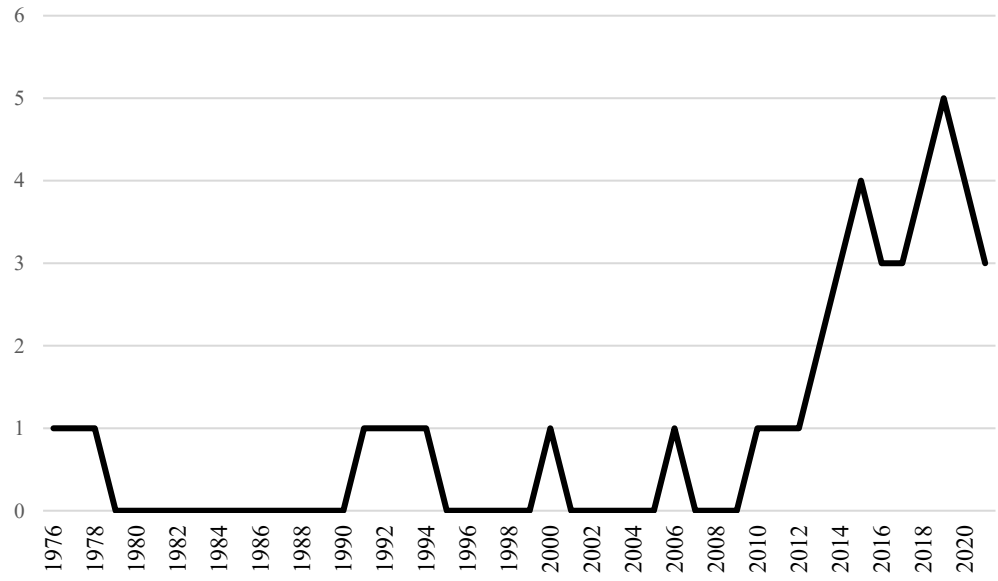

Figure 3: Evolution in the Number of Publications on the Use of BARS for the Assessment of Teaching Effectiveness

These publications received a total of 451 citations, with an average citation of 13.23. A h-index of 11 is also observed, indicating that at least 11 of these works received 11 citations or more.

Regarding the thematic categories, the papers on this issue were published in journals concentrated in education and psychology categories, such as educational research, education special, educational psychology, and psychology applied.

Considering the citations obtained (from the highest number to the lowest number), the most significant papers in this area include studies by Murphy and Anhalt (1992), Martin-Raugh et al. (2016), Matosas-López, Aguado-Franco, and Gómez-Galán (2019), Arnăutua and Panc (2015), and Matosas-López, Romero-Ania, and Cuevas-Molano (2019).

Table 1: High-Impact Publications on the Use of BARS for the Assessment of Teaching Effectiveness

Author(s)	*Educational System Stage*	*Research Approach*	*Research Methodology*	*Techniques Used*
Murphy and Anhalt (1992)	University	Quantitative	Psychometric research	Questionnaires
Martin-Raugh et al. (2016)	Elementary school	Quantitative and qualitative	Comparative/experimental research	Video-taped lessons and group discussions
Matosas-López, Aguado-Franco, and Gómez-Galán (2019)	University	Quantitative and qualitative	Instrumental research	Panel of experts, group interviews, and questionnaires
Arnăutua and Panc (2015)	University	Qualitative	Instrumental research	Panel of experts
Matosas-López, Romero-Ania, and Cuevas-Molano (2019)	University	Quantitative	Comparative/experimental research	Questionnaires

Murphy and Anhalt (1992), in a laboratory study, explore the halo error on these scales analyzing specifically student and teacher behavior. The authors address that the halo error on the examined instrument depends basically on the teaching situation and teacher behavior. When teacher performance is viewed by the student as constant, the halo error is moderately stable, while when teacher behavior is changing, halo errors are highly unstable.

The study by Martin-Raugh et al. (2016), using video-taped teacher lessons, interviews, and surveys, compare a BARS instrument designed to measure teaching quality in primary school with a traditional instrument. The findings indicate that the BARS instrument examined is better in aspects such as its perceived ease of use or accuracy. In this study, the authors conclude that appraisal instruments that use well-defined behavioral examples as points on the scales allow a much more complete and adequate measurement of teaching effectiveness.

The research by Matosas-López, Aguado-Franco, and Gómez-Galán (2019) describes in detail the procedure to construct a BARS questionnaire to assess teaching quality in blended learning modalities in the university setting. The authors, following the traditional guidelines for the design of BARS, create an instrument that facilitates the appraisal of critical aspects from distance learning modalities such as teacher–student communication, the use of advanced digital learning resources, the course design, or the teacher's technical competencies. The authors conclude that the final instrument provides a clear and unambiguous feedback that not only enables the teacher to take specific corrective measures, but also reinforces the formative purpose of the assessment.

Arnăutua and Panc (2015), following the traditional guidelines for the construction of BARS questionnaires, propose the use of behavioral scales to evaluate the performance of faculty members. The authors address that the resulting instrument is not only valid for the appraisal of teaching activity but is a system that provides a 360-degrees feedback ensuring objectivity in the assessment and offering the teacher the possibility of adjusting their own performance through others' perspective.

Finally, the study by Matosas-López, Romero-Ania, and Cuevas-Molano (2019) analyzes the completion time in teacher evaluation surveys that applied participation incentives, comparing BARS and Likert questionnaires. The findings reveal that completion times, when participation incentives are applied, vary depending on the type of questionnaire used. While Likert scales do not favor the correct reading and completion of surveys, this situation can be improved by using BARS questionnaires. According to the authors, the use of BARS, with examples of behavior that

are familiar to the student, motivates them to do a more detailed reading of the questionnaire, investing the time necessary to give satisfactory answers to the survey questions.

Discussion and Conclusion

The measurement of teaching effectiveness has been widely analyzed. However, traditionally, these measurement mechanisms have been put into practice through surveys with Likert-type questionnaires. The persistent use of this assessment pattern can be seen in the works by Remmers (1928, 1971) in the 1920s, Marsh in the 1980s and 1990s (1982, 1991), and, more recently, by Spooren (2010). Despite its popularity, different studies have shown that this measurement system has notable limitations (Hornstein 2017; Spooren, Brockx, and Mortelmans 2013).

The BARS, according to various authors, makes it possible to overcome many of the limitations of traditional questionnaires. The instrument suffers from fewer halo effects and lenience errors (Bernardin, Alvares, and Cranny 1976; Borman and Dunnette 1975; Campbell, Dunnette, and Arvey 1973), improves validity, and reduces the influence of bias (Debnath, Lee, and Tandon 2015; Martin-Raugh et al. 2016; Ohland, Loughry, and Woehr 2012), and it provides interrater reliability indicators that are moderately higher than those found in other instruments (Bearden and Wagner 1988; Debnath, Lee, and Tandon 2015; Williams and Seiler 1973).

Despite the above, the use of BARS questionnaires is not exempt from drawbacks. Their main shortcomings are the heavy investment of time required for instrument construction (Goodale and Burke 1975; Stoskopf et al. 2016), the difficulty in obtaining behavioral examples to represent scale midpoints (Debnath, Lee, and Tandon 2015; Hauenstein, Brown, and Sinclair 2010), and the loss of behavioral information throughout the construction process (Matosas-López, Leguey-Galán, and Leguey-Galán 2019; Schwab, Heneman, and DeCotiis 1975). Such issues discourage the use of BARS in many cases.

The literature review conducted by the author reveals that, although the first publications date from the 1960s and 1970s, BARS did not become widespread until recent years. The present study also shows that in addition to the high acceptance of these scales in the healthcare context for the evaluation of disorders and sicknesses, their use in the educational setting for the measurement of teaching effectiveness is also confirmed.

In this educational context, the first publications on BARS are comparative studies on the halo effect, leniency error, and interrater reliability carried out by Bernardin, Alvares, and Cranny (1976) with university professors. However, even though the rise of BARS occurred in the 1970s, the use of this type of scale for the measurement of teaching performance did not gain popularity until the beginning of 2010. Studies such as those by Arnăutua and Panc (2015) on the evaluation of university professors and by Martin-Raugh et al. (2016) on measuring the efficiency of elementary education teachers are key in this sense.

In the author's opinion, the increase in the number of publications on the use of BARS for the evaluation of teaching effectiveness is conditioned by the positive impact of information and communication technologies (ICTs) in the arduous process of instrument construction. As has been previously pointed out, the high investment of time required for BARS design can discourage its use. The construction process, in accordance with the standards, necessitates the use of panels of experts, group interviews, and surveys. Nevertheless, recent research indicates that ICTs can substantially reduce the time investment needed. Thus, for example, video conference tools facilitate both the panels of expert meetings required to describe the teaching dimensions and the group interviews needed to gather critical incidents (Matosas-López, Aguado-Franco, and Gómez-Galán 2019). Similarly, the use of web forms facilitates the online completion of the surveys required to determine the degree of importance given by the evaluators to the different critical incidents that serve as anchor points in the final scale (Matosas-López, Leguey-Galán, and Doncel-Pedrera 2019).

Another aspect that, in the author's opinion, has positively influenced the increase in the use of BARS is the awareness of the importance of obtaining measurements more aligned with the reality of teacher performance. In the context of educational organization, these assessments are directly considered by quality agencies, public administrations, and private organizations when making decisions about teacher promotion (Ibáñez-López, Hernández-Pina, and Monroy 2020). Therefore, these evaluations can neither be taken lightly nor conducted with inaccurate mechanisms. Given such imperatives, recent studies postulate the use of BARS as an alternative measurement system (Matosas-López et al. 2021).

According to this author, the positive incidence of ICTs in the design of the instrument, in addition to the growing awareness of the importance of these evaluations for the professional promotion of teachers, has contributed to the recent increase in the number of publications on this topic.

The present study reveals an evident interest of the academic community in the use of BARS, in the educational setting, for the assessment of teaching effectiveness. However, the use of these questionnaires continues to be low when compared with other measurement systems, and perhaps also less efficient. In view of the above, the author concludes that research on the use of BARS for the evaluation of teacher performance is not only an emerging line of study, but also a topic that invites the academic community to reflect on which mechanisms are most appropriate when dealing with a critical task, in the context of educational organization, such as the appraisal of teaching effectiveness.

Limitations and Further Research

This work is not without limitations. In this sense, despite the fact that this literature review considers the catalogs in the WOS Main Collection, this decision leaves out publications on the topic, for example, that are indexed only in Scimago Journal Rank database of SCOPUS, or Education Resources Information Center (ERIC), as well as publications that do not appear in WOS.

In the same way, the author retrieves only those publications under the formats of journal articles, review articles, and book chapters. However, future studies could also consider works of lesser visibility in the form of conference proceedings, meeting abstracts, notes, reports, theses, or working papers. Examples of works with a certain soundness with publication formats such as those previously mentioned, and which have therefore been left out of this literature review, are those of Kell et al. (2017), Klieger et al. (2018), and Coyne (2020), among others.

Finally, the present literature review does not explore the potential connections of the topic of teaching effectiveness measurement with other theories such as the Bloom's Taxonomy Theory (Krathwohl 2002), the Jerome Bruner's Spiral Curricular Theory (Takaya 2008), or the Theories of the Constructivism and Behaviorism Teaching Methodologies (Almala 2005).

Therefore, this work, besides inviting the academic community to reflect on which instruments are most suitable to assess teaching effectiveness, also addresses as a future line of research the possibility of conducting new reviews on the topic considering other secondary databases and publication formats.

REFERENCES

Almala, Abed H. 2005. "A Constructivist Conceptual Framework for a Quality e-Learning Environment." *Distance Learning* 2 (5): 9. https://www.proquest.com/openview/4ad504cfd0df08b49156b35d3efddb1f/1?pq-origsite=gscholar&cbl=29704.

Alvarado-Lagunas, Elías, Dionisio Ramírez, and Ernesto Téllez. 2016. "Percepción de la calidad educativa: caso aplicado a estudiantes de la Universidad Autónoma de Nuevo León y del Instituto Tecnológico de Estudios Superiores de Monterrey" [Perception of Educational Quality: Case Applied to Students of the Autonomous University of Nuevo León and the Technological Institute of Higher Studies of Monterrey]. *Revista de la Educacion Superior* [Higher Education Journal] 45 (180): 55–74. https://doi.org/10.1016/j.resu.2016.06.006.

Arnăutua, Elena, and Ioana Panc. 2015. "Evaluation Criteria for Performance Appraisal of Faculty Members." *Procedia—Social and Behavioral Sciences* 203:386–392. https://doi.org/10.1016/j.sbspro.2015.08.313.

Arvey, Richard, and Joseph Hoyle. 1974. "A Guttman Approach to the Development of Behaviorally Based Rating Scales for Systems Analysts and Programmer/Analysts." *Journal of Applied Psychology* 59 (1): 61–68. https://doi.org/10.1037/h0035830.

Barnokhon, Negmatova. 2022. "Differences between Assessment and Evaluation." *Journal of Pedagogical Inventions and Practices* 8:41–43. https://zienjournals.com/index.php/jpip/article/view/1572.

Bearden, Ronald, and Michael Wagner. 1988. "Developing Behaviorally Anchored Rating Scales for the Machinist's Mate Rating." https://apps.dtic.mil/dtic/tr/fulltext/u2/a195403.pdf.

Bernardin, John. 1977. "Behavioral Expectation Scales versus Summated Scales." *Journal of Applied Psychology* 62 (4): 422–427. http://psycnet.apa.org/record/1978-09104-001.

Bernardin, John, Kenneth Alvares, and Charles Cranny. 1976. "A Recomparison of Behavioral Expectation Scales to Summated Scales." *Journal of Applied Psychology* 61 (5): 564–570. https://doi.org/10.1037/0021-9010.61.5.564.

Bernardin, John, and Richard Beatty. 1984. *Performance Appraisal: Assessing Human Behavior at Work*. Boston: Kent.

Bernardin, John, and Patricia Smith. 1981. "A Clarification of Some Issues Regarding the Development and Use of Behaviorally Anchored Ratings Scales (BARS)." *Journal of Applied Psychology* 66 (4): 458–463. https://doi.org/10.1037/0021-9010.66.4.458.

Borman, Walter. 1991. "Job Behavior, Performance, and Effectiveness." In *Handbook of Industrial and Organizational Psychology*, edited by Marvin Dunnette, 271–326. Palo Alto: Consulting Psychologists Press. Borman, Walter, and Marvin Dunnette. 1975. "Behavior-Based versus Trait-Oriented Performance Ratings: An Empirical Study." *Journal of Applied Psychology* 60 (5): 561–565. https://doi.org/10.1037/0021-9010.60.5.561.

Campbell, John, Marvin Dunnette, and Richard Arvey. 1973. "The Development and Evaluation of Behaviorally Based Rating Scales." *Journal of Applied Psychology* 57 (1): 15–22. https://doi.org/10.1037/h0034185.

Carretta, Thomas, and Laurie Walters. 1991. *The Development of Behaviorally Anchored Rating Scales (BARS) for Evaluating USAF Pilot Training Performance*. Arlington, TX: Armstrong Laboratory, Manpower and Personnel Research Division.

Catano, Victor. 2007. "Performance Appraisal of Behavior-Based Competencies: A Reliable and Valid Procedure." *Personnel Psychology* 60:201–230. https://doi.org/10.1111/j.1744-6570.2007.00070.x.

Coyne, Bryan. 2020. "Western Kentucky University Psychological Sciences Faculty Bars Revision." Masters thesis, Western Kentucky University. https://digitalcommons.wku.edu/theses.

Debnath, Sukumar, Brian B. Lee, and Sudhir Tandon. 2015. "Fifty Years and Going Strong: What Makes Behaviorally Anchored Rating Scales so Perennial as an Appraisal Method?" *International Journal of Business and Social Science* 6 (2): 16–25. https://www.semanticscholar.org/paper/Fifty-Years-and-Going-Strong%3A-What-Makes-Anchored-Debnath-Lee/9c381f049faa17915d904472fa09f67ed6b2ec46.

Fernández-Millán, Juan, and Marina Fernández-Navas. 2013. "Development of an Evaluation Performance Scale for Social Educators in Child Protection Centers." *Intangible Capital* 9 (3): 571–589. https://doi.org/10.3926/ic.410.

Flanagan, Jhon. 1954. "The Critical Incident Technique." *Psychological Bulletin* 51 (4): 327–358. https://doi.org/10.1037/h0061470.

Fogli, Laurence, Charles Hulin, and Milton Blood. 1971. "Development of First-Level Behavioral Job Criteria." *Journal of Applied Psychology* 55 (1): 3–8. https://doi.org/10.1037/h0030631.

Gómez-Gómez, Marta, and Lorenzo García-Aretio. 2016. "La formación como factor clave en la integración de la Pizarra Digital Interactiva. Perspectivas de profesores y coordinadores TIC" [Training as a Key Factor in the Integration of the Interactive Whiteboard. Perspectives of Teachers and ICT Coordinators]. *Revista Electrónica Interuniversitaria de Formación del Profesorado* [Interuniversity Electronic Journal of Teacher Training] 19 (3): 35. https://doi.org/10.6018/reifop.19.3.225451.

Goodale, James, and Ronald Burke. 1975. "Behaviorally Based Rating Scales Need Not Be Job Specific." *Journal of Applied Psychology* 60 (3): 389–391. https://doi.org/10.1037/h0076629.

Hauenstein, Neil, Reagen Brown, and Andrea Sinclair. 2010. "BARS and Those Mysterious, Missing Middle Anchors." *Journal of Business and Psychology* 25 (4): 663–672. https://doi.org/10.1007/s10869-010-9180-7.

Hernandez-Ramos, Juan, and Fernando Martínez-Abad. 2021. "La importancia de la actitud del docente universitario: validación de una escala para su consideración" [The Importance of the Attitude of the University Teacher: Validation of a Scale for Its Consideration]. *Revista Electrónica Interuniversitaria de Formación del Profesorado* [Interuniversity Electronic Journal of Teacher Training] 24 (1): 59–71. https://doi.org/10.6018/REIFOP.414781.

Hom, Peter, Angelo DeNisi, Angelo Kinicki, and Brendan Bannister. 1982. "Effectiveness of Performance Feedback from Behaviorally Anchored Rating Scales." *Journal of Applied Psychology* 67 (5): 568–576. https://doi.org/10.1037/0021-9010.67.5.568.

Hornstein, Henry. 2017. "Student Evaluations of Teaching Are an Inadequate Assessment Tool for Evaluating Faculty Performance." *Cogent Education* 4 (1): 1–8. https://doi.org/10.1080/2331186X.2017.1304016.

Hoy, Wayne, and Anita Woolfolk. 1993. "Teachers' Sense of Efficacy and the Organizational Health of Schools." *Elementary School Journal* 93 (4): 355–372. https://doi.org/10.1086/461729.

Ibáñez-López, Francisco, Fuensanta Hernández-Pina, and Fuensanta Monroy. 2020. "Evaluación y acreditación de titulaciones universitarias en Educación desde el punto de vista del profesorado" [Evaluation and Accreditation of University Degrees in Education from the Point of View of Teachers]. *Revista Interuniversitaria de Formación del Profesorado* [Interuniversity Electronic Journal of Teacher Training] 34 (3): 137–154. https://doi.org/10.47553/RIFOP.V34I3.81380.

Jacobs, Rick, Ditsa Kafry, and Sheldon Zedeck. 1980. "Expectations of Behaviorally Anchored Rating Scales." *Personnel Psychology* 33 (3): 595–640. https://doi.org/10.1111/j.1744-6570.1980.tb00486.x.

Kavanagh, Michael, and John Duffy. 1978. "An Extension and Field Test of the Retranslation Method for Developing Rating Scales." *Personnel Psychology* 31 (3): 461–470. https://doi.org/10.1111/j.1744-6570.1978.tb00455.x.

Kell, Harrison, Michelle Martin-Raugh, Lauren Carney, Patricia Inglese, Lei Chen, and Gary Feng. 2017. "Exploring Methods for Developing Behaviorally Anchored Rating Scales for Evaluating Structured Interview Performance." https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12152.

Kitsantas, Anastasia. 2012. "Teacher Efficacy Scale for Classroom Diversity (TESCD): A Validation Study." *Profesorado: Revista de Curriculum y Formación del Profesorado* [Faculty: Curriculum Journal and Faculty Formation] 16 (1): 35–44. https://doi.org/http://hdl.handle.net/10481/23014.

Klieger, David, Harrison Kell, Samuel Rikoon, Kri Burkander, Jennifer Bochenek, and Jane Shore. 2018. "Development of the Behaviorally Anchored Rating Scales for the Skills Demonstration and Progression Guide." https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12210.

Krathwohl, David. 2002. "A Revision of Bloom's Taxonomy: An Overview." *Theory into Practice* 41 (4): 212–218. https://doi.org/10.1207/s15430421tip4104_2.

Landy, Frank, and James Guion. 1970. "Development of Scales for the Measurement of Work Motivation." *Organizational Behavior and Human Performance* 5 (1): 93–103. https://doi.org/10.1016/0030-5073(70)90007-3.

Langton, Robert. 2013. "Designing a Quality Assurance System for an Australian Business School." *International Journal of Educational Organization and Leadership* 19 (1): 41–46. https://doi.org/10.18848/2329-1656/CGP/V19I01/48517.

Leguey Galán, Santiago, Sonsoles Leguey Galán, and Luis Matosas López. 2018. "¿De qué depende la satisfacción del alumnado con la actividad docente?" [What Drives Students' Satisfaction with Instruction?] *Revista Espacios* 39 (17): 13–29. http://www.revistaespacios.com/a18v39n17/18391713.html.

López-Cámara, Ana, Ignacio González-López, and Carlota De Leon-Huertas. 2014. "Perfil de un buen docente. Aplicación de un protocolo de evaluación de las competencias del profesorado universitario" [Profile of a Good Teacher. Application of a Protocol for Evaluating the Competences of University Teaching Staff]. *Revista Electrónica Interuniversitaria de Formación del Profesorado* [Interuniversity Electronic Journal of Teacher Training] 17 (1): 133–148. https://doi.org/10.6018/REIFOP.17.1.190531.

Lynott, Donna, and Anita Woolfolk. 1994. "Teachers' Implicit Theories of Intelligence and Their Educational Goals." *Journal of Research & Development in Education* 27 (4): 253–264. https://psycnet.apa.org/record/1995-15122-001.

MacDonald, Heather, and Lorne Sulsky. 2009. "Rating Formats and Rater Training Redux: A Context-Specific Approach for Enhancing the Effectiveness of Performance Management." *Revue canadienne des sciences du comportement* [Canadian Journal of Behavioural Science] 41 (4): 227–240. https://doi.org/10.1037/a0015165.

Marsh, Herbert. 1982. "SEEQ: A Reliable, Valid, and Useful Instrument for Collecting Students' Evaluations of University Teaching." *British Journal of Educational Psychology* 52 (2): 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x.

Marsh, Herbert. 1991. "A multidimensional Perspective on Students Evaluations of Teaching Effectiveness—Reply to Abrami and Dapollonia (1991)." *Journal of Educational Psychology* 83 (3): 416–421. https://doi.org/10.1037//0022-0663.83.3.416.

Marsh, Herbert, and Dennis Hocevar. 1991. "The Multidimensionality of Students' Evaluations of Teaching Effectiveness: The Generality of Factor Structures across Academic Discipline, Instructor Level, and Course Level." *Teaching and Teacher Education* 7 (1): 9–18. https://doi.org/10.1016/0742-051X(91)90054-S.

Martin-Raugh, Michelle, Richard Tannenbaum, Cynthia Tocci, and Clyde Reese. 2016. "Behaviorally Anchored Rating Scales: An Application for Evaluating Teaching Practice." *Teaching and Teacher Education* 59:414–419. https://doi.org/10.1016/j.tate.2016.07.026.

Matosas, Luis. 2018. "Aspectos de comportamiento básico del profesor universitario en los procesos de valoración docente para modalidades blended learning" [Core Behavioral Aspects of University Professors in Student Evaluation of Teaching (SET) for Blended Learning Models]. *Revista Espacios* 39 (10): 10–24. https://www.revistaespacios.com/a18v39n10/18391010.html.

Matosas-López, Luis, Juan Carlos Aguado-Franco, and José Gómez-Galán. 2019. "Construcción de un instrumento con escalas de comportamiento para la evaluación la calidad docente en modalidades blended learning" [Constructing an Instrument with Behavioral Scales to Assess Teaching Quality in Blended Learning Modalities]. *Journal of New Approaches in Educational Research* 8 (2): 142–165. https://doi.org/10.7821/naer.2019.7.410.

Matosas-López, Luis, and Cesar Bernal-Bravo. 2020. "Presencia de las TIC en el diseño de un instrumento BARS para la valoración de la eficiencia del profesorado en modalidades de enseñanza online" [Presence of ICT in the Design of BARS for the Assessment of Teaching Efficiency in Online Modalities]. *Psychology, Society, & Education* 12 (1): 43–56. https://doi.org/10.25115/psye.v0i0.2501.

Matosas-López, Luis, Cesar Bernal-Bravo, Alberto Romero-Ania, and Irene Palomero-Ilardia. 2019. "Quality Control Systems in Higher Education Supported by the Use of Mobile Messaging Services." *Sustainability* 11 (21): 6063. https://doi.org/10.3390/su11216063.

Matosas-López, Luis, and Beatriz García-Sánchez. 2019. "Beneficios de la distribución de cuestionarios web de valoración docente a través de mensajería SMS en el ámbito universitario: tasas de participación, inversión de tiempo al completar el cuestionario y plazos de recogida de datos" [Benefits in the Distribution of Evaluation of Teaching Web Questionnaires through SMS Messaging in the University Context: Participation Rates, Investment of Time when Completing the Questionnaire and Data Collection Periods]. *Revista Complutense de Educación* [Complutense Education Journal] 30 (3): 831–845. https://doi.org/10.5209/RCED.59224.

Matosas-López, Luis, Santiago Leguey-Galán, and Luis Miguel Doncel-Pedrera. 2019. "Converting Likert Scales into Behavioral Anchored Rating Scales(Bars) for the Evaluation of Teaching Effectiveness for Formative Purposes." *Journal of University Teaching & Learning Practice* 16 (3): 1–24. https://doi.org/10.53761/1.16.3.9.

Matosas-López, Luis, Santiago Leguey-Galán, and Sonsoles Leguey-Galán. 2019. "Cómo resolver el problema de pérdida de información conductual en el diseño de Behaviorally Anchored Rating Scales-BARS. El caso de la medición de la eficiencia docente en el contexto universitario" [How to Deal with the Problem of Loss of Behavioral Information during the Construction of Behaviorally Anchored Rating Scales (BARS): The Teaching Efficiency Assessment Case in the University Context]. *Revista Espacios* 40 (19): 6–21. http://www.revistaespacios.com/a19v40n19/19401906.html.

Matosas-López, Luis, Alberto Romero-Ania, and Elena Cuevas-Molano. 2019. "¿Leen los Universitarios las Encuestas de Evaluación del Profesorado Cuando se Aplican Incentivos por Participación? Una Aproximación Empírica" [Do Students Read Teacher Evaluation Surveys when Participation Incentives Are Applied? An Empirical Approach]. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación* [Ibero-American Journal on Quality, Effectiveness and Change in Education] 17 (3): 99–124. https://doi.org/10.15366/reice2019.17.3.006.

Matosas-López, Luis, Roberto Soto-Varela, Melchor Gómez-García, and Moussa Boumadan. 2021. "Quality Systems for a Responsible Management in the University." In *Sustainable and Responsible Entrepreneurship and Key Drivers of Performance*, edited by Cristina Popescu and Rahul Verma, 33–58. Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-7998-7951-0.CH006.

Medley, Donald. 1977. *Teacher Competence and Teacher Effectiveness. A Review of Process-Product Research*. Washington, DC: Order Department, American Association of Colleges for Teacher Education.

Mula-Falcón, Javier. 2021. "How Does Performance Evaluation Impact Academics?" *International Journal of Assessment and Evaluation* 29 (1): 27–36. https://doi.org/10.18848/2327-7920/CGP/V29I01/27-36.

Murphy, Kevin, and Rebecca Anhalt. 1992. "Is Halo Error a Property of the Rater, Ratees, or the Specific Behaviors Observed?" *Journal of Applied Psychology* 77 (4): 494–500. https://doi.org/10.1037/0021-9010.77.4.494.

Murphy, Kevin, and Virginia Pardaffy. 1989. "Bias in Behaviorally Anchored Rating Scales: Global or Scale-Specific?" *Journal of Applied Psychology* 74 (2): 343–346. https://doi.org/10.1037/0021-9010.74.2.343.

Ohland, Mathew, Misty Loughry, and David Woehr. 2012. "The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation." *Academy of Management Learning & Education* 11 (4): 609–630. https://doi.org/10.5465/amle.2010.0177.

Remmers, Herbert. 1928. "The Relationship between Students' Marks and Student Attitude toward Instructors." *School & Society* 28:759–760. http://psycnet.apa.org/psycinfo/1929-01345-001.

Remmers, Herbert. 1971. "Rating Methods in Research of Teaching." In *Handbook of Research on Teaching*, edited by Neil Gage, 184–201. Chicago: Rand McNally.

Reynolds, David, and Charles Teddlie. 2000. *The International Handbook of School Effectiveness Research*. London: Routledge.

Ruiz-Ariza, Alberto, and Cristina Cruz-González. 2021. "How to Design Systematic Reviews of the Literature in Educational Sciences?" *International Journal of Design Education* 15 (1): 165–175. https://doi.org/10.18848/2325-128X/CGP/V15I01/165-175.

Schwab, Donald, Ian Heneman, and Thomas DeCotiis. 1975. "Behaviorally Anchored Rating Scales: A Review of the Literature." *Personnel Psychology* 28 (4): 549–562.

Sharon, Amiel, and Carl Bartlett. 1969. "Effect of Instructional Conditions in Producing Leniency on Two Types of Rating Scales." *Personnel Psychology* 22 (3): 251–263. https://doi.org/10.1111/j.1744-6570.1969.tb00330.x.

Smith, Patricia, and Louise Kendall. 1963. "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales." *Journal of Applied Psychology* 47 (2): 149–155. https://doi.org/10.1037/h0047060.

Spooren, Pieter. 2010. "On the Credibility of the Judge: A Cross-Classified Multilevel Analysis on Students' Evaluation of Teaching." *Studies in Educational Evaluation* 36 (4): 121–131. https://doi.org/10.1016/j.stueduc.2011.02.001.

Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83 (4): 598–642. https://doi.org/10.3102/0034654313496870.

Spooren, Pieter, and Wim Christiaens. 2017. "I Liked Your Course because I Believe in (the Power of) Student Evaluations of Teaching (SET). Students' Perceptions of a Teaching Evaluation Process and Their Relationships with SET Scores." *Studies in Educational Evaluation* 54:43–49. https://doi.org/10.1016/j.stueduc.2016.12.003.

Stoskopf, Carlin H., Deborah C. Glik, Samuel L. Baker, James R. Ciesla, and Catherine M. Cover. 2016. "The Reliability and Construct Validity of a Behaviorally Anchored Rating Scale Used to Measure Nursing Assistant Performance." *Evaluation Review* 16 (3): 333–345. https://doi.org/10.1177/0193841X9201600307.

Takaya, Keiichi. 2008. "Jerome Bruner's Theory of Education: From Early Bruner to Later Bruner." *Interchange* 39 (1): 1–19. https://doi.org/10.1007/S10780-008-9039-2.

Vanacore, Amalia, and María Pellegrino. 2019. "How Reliable Are Students' Evaluations of Teaching (SETs)? A Study to Test Student's Reproducibility and Repeatability." *Social Indicators Research* 54 (2): 241–260. https://doi.org/10.1007/s11205-018-02055-y.

Veciana-Vergés, José, and Joan Capelleras-i-Segura. 2004. "Calidad de servicio en la enseñanza universitaria desarrollo y validación de una escala media" [Quality of Service in University Education Development and Validation of an Average Scale]. *Revista Europea de Dirección y Economía de la Empresa* [European Journal of Business Management and Economics] 13 (4): 55–72. https://dialnet.unirioja.es/servlet/articulo?codigo=1121821.

Wea, Donauts, and Basilius Werang. 2020. "Teachers' Working Conditions and Job Performance in the Elementary Schools of Indonesia: A Survey from Southern Papua." *International Journal of Educational Organization and Leadership* 27 (1): 37–46. https://doi.org/10.18848/2329-1656/CGP/V27I01/37-46.

Williams, Willian, and Dale Seiler. 1973. "Relationship between Measures of Effort and Job Performance." *Journal of Applied Psychology* 57 (1): 49–54. https://doi.org/10.1037/h0034201.

Woods, Robert, Michael Sciarini, and Deborah Breiter. 1998. "Performance Appraisals in Hotels: Widespread and Valuable." *Cornell Hotel and Restaurant Administration Quarterly* 39 (2): 25–29. https://doi.org/10.1177/001088049803900205.

Zhao, Jing, and Dorinda Gallant. 2012. "Student Evaluation of Instruction in Higher Education: Exploring Issues of Validity and Reliability." *Assessment & Evaluation in Higher Education* 37 (2): 227–235. https://doi.org/10.1080/02602938.2010.523819.

ABOUT THE AUTHOR

Luis Matosas-López: Associate Professor, Department of Financial Economics and Accounting, Rey Juan Carlos University, Madrid, Spain