The Future of Engineering Education
2024 Annual Conference & Exposition
Oregon Convention Center
Portland, OR . June 23 - 26, 2024
ASEE
Paper ID #43738

Enhancing Teaching Evaluation in Engineering Education: An Exploration of the Theory of Rating

Mayar Madboly, Virginia Polytechnic Institute and State University

Mayar Madboly is currently a PhD student in the department of Engineering Education at Virginia Polytechnic Institute and University. Her research focuses on the issues in teaching evaluation, teaching effectiveness, and teamwork dynamics in engineering student and practitioner teams. She received her Bachelor's and Master's degrees in Electrical Energy Engineering from the faculty of Engineering in Cairo University in Egypt.

Dr. Nicole P. Pitterson, Virginia Polytechnic Institute and State University

Nicole is an assistant professor in the Department of Engineering Education at Virginia Tech. Prior to joining VT, Dr. Pitterson was a postdoctoral scholar at Oregon State University. She holds a PhD in Engineering Education from Purdue University and oth

Enhancing Teaching Evaluation in Engineering Education: An Exploration of the Theory of Rating

Abstract

Teaching evaluation in higher education is an essential practice that plays a pivotal role in ensuring the quality and effectiveness of academic instruction. It involves the systematic assessment of teaching methods, strategies, and their outcomes, allowing institutions to gauge the overall performance of educators and identify areas for improvement. This process allows educators to reflect on their teaching practices, adapt to evolving pedagogical trends, and enhance their students' learning experiences. In the existing literature much is known about how teaching evaluations are conducted and their value in helping educators become better at their craft. However, there remains a gap in our understanding of the theoretical underpinnings of how supervisors and peer evaluators make decisions about how to rate teaching beyond their own perceptions of teaching.

In this paper, we introduce the theory of rating (ToR) by Robert Wherry as a candidate theoretical framework for studying teaching evaluation. The ToR explains sources of error and bias in ratings and methods to minimize their impact. The ToR also demonstrates important aspects of rating scales and settings and talks about methods used to test rating reliability and control bias. Although the ToR was developed in 1952 to account for all dimensions of rating/evaluation, it is not yet popular in studying teaching evaluation. Thus, we aim in this paper to widen our understanding of teaching evaluation dimensions by introducing and explaining the ToR along with its hypotheses then show how the theory was applied in previous literature. Most importantly, we will show the adequacy of this theory to study teaching evaluation and suggest steps to improve the teaching evaluation process. Also, we will compare the theory principles to current standards of teaching evaluation.

Introduction

In 1952 Robert J. Wherry developed the theory of rating (ToR), the theory was republished in 1982 by Christopher J. Barlett with some minor editing to make the equations more readable and the assumptions more understandable [1]. The ToR consists of 46 theorems which appear in equation form and tackles varied constructs (see appendix I for examples), most of the constructs have at least two hypotheses (corollaries) to show nuances between the constructs [1].

The ToR studies ratings, also called evaluations of performance, suggests ways to minimize bias and error in ratings, sets the main guidelines for designing rating scales and settings, and explains different methods of testing the reliability of ratings and controlling bias in responses. Our goal is to highlight the foundations of a valid evaluation system that can serve as a tool in teaching evaluation. The ToR was used a lot to study managerial and organizational leadership [2], [3], [4], and [5] and was used to a much lesser extent in educational settings. Some of the educational applications of the ToR include teaching observations [6], and grading students [7]. Due to the afore-mentioned reasons, the objective of this paper is fourfold: explain the ToR, show its existing applications, highlight ways in which the ToR can be used as a theoretical framework to study teaching evaluations, and compare current teaching evaluation standards to the ToR principals.

Theory overview

As described by the ToR [8], the rating scene entails “a rater attempts to make a report upon the past behavior of a ratee in some special area defined by a rating item”. Usually, ratings cover a specific period, which is why accurate observation and accurate recall of observation are necessary to deliver accurate rating responses. Thus, factors which influence rating results are threefold; “performance of the ratee, observation of performance by the rater, recall of observation by the rater” [1]. The performance of the ratees has three components: true ability, error, and environmental factors. Also, Observation has three components: observed performance, bias, and error. Similarly, recall of observation is made up of remembered observations, error in remembering, and bias in remembering. Ultimately, the ToR aims to maximize the weight of true ability in the rating result and minimize the weights of other components like error in perception, error in recall, bias in perception, bias in recall and environmental factors. It is worth noting that this theory deals with implicit bias and neglects cases where intentional falsification of rating responses may take place. Also, the theory uses some gendered terminology like “man-paced” which reflects use of language during that time.

The ToR consists of 46 theorems and most of the theorems have a set of corollaries (hypotheses). In this paper we will discuss the first 27 theorems and their associated corollaries because they are most aligned with the design aspects and settings of teaching evaluation, the first 27 theorems show necessary steps taken before and during an evaluation to secure accurate results. However, theorems 28 to 46 deal with interpreting a rating result, detecting and removing bias from evaluation responses which means that these theorems focus on after-evaluation actions. Based on the definitions and focus of the 27 theorems they were clustered into six major constructs. Figure 1 shows the main constructs of the theory: ratees’ control and their relationship with raters, rating settings, rating design, knowledge, training, and bias, bias control and the final major construct is testing the reliability of rating response. Figure 1 also shows the application of each major construct. Applications vary from minimizing error and bias in evaluation systems, designing evaluation systems and evaluation settings. Moreover, the theory sets the guidelines on how to test the reliability of evaluations and how to remove bias and error components from the evaluation results. The following sections will refer to and discuss the theory constructs, hypotheses, theory applications, how to apply the theory in teaching evaluation and how the theory principles compare to the current teaching evaluation standards.

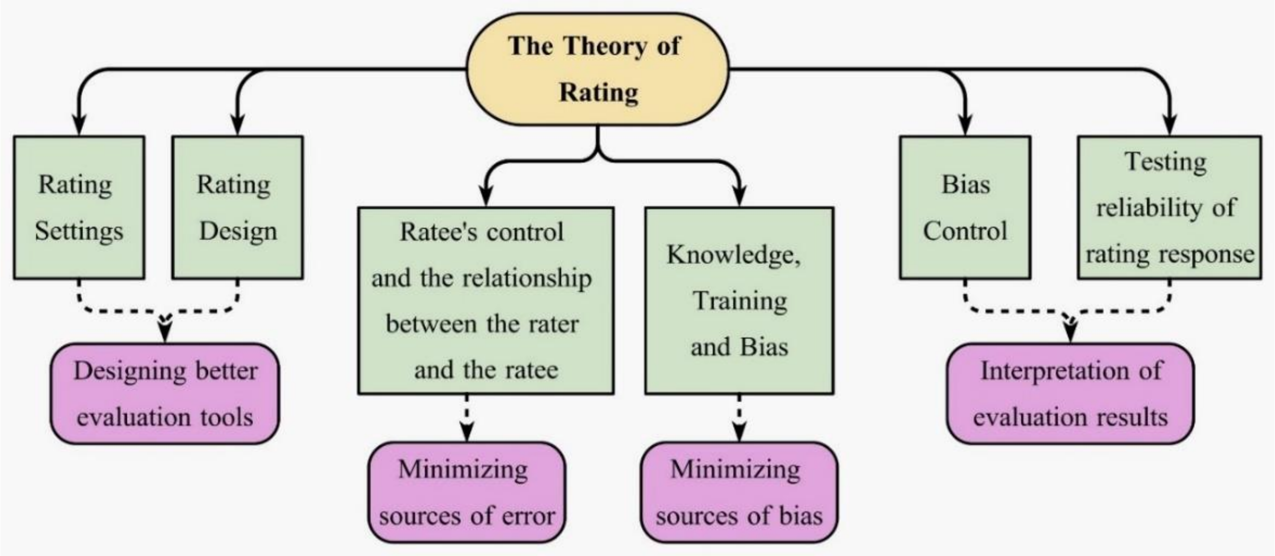
The Theory of Rating
Rating Settings
Rating Design
Ratee's control and the relationship between the rater and the ratee
Knowledge, Training and Bias
Bias Control
Testing reliability of rating response
Designing better evaluation tools
Minimizing sources of error
Minimizing sources of bias
Interpretation of evaluation results

Figure 1: ToR constructs and applications

Ratees' control and their relationship with raters

This major theory construct is covered by theorems 1, 3 and 4 and hypotheses 1.a, 1.b, 1.c, 3.a, and corollaries 4.a, 4.b, and 4.c as shown in table I below.

Table I: ToR theorems and corollaries ratee's control and their relationship with raters

Theorem No.	Definition	Associated Corollaries
1	"Tasks in which the performance is maximally controlled by the ratee rather than by the work situation will be more favorable to accurate ratings" [1].	1a. "Tasks in which the raw material, tools, working conditions (light, heat, etc.) are constant from worker to worker will lead to more accurate ratings of ability than will those in which such factors are variable" [1]. 1b. "Tasks which are man-paced rather than machine-paced will lead to more accurate ratings of ability" [1]. 1c. "Positions in which output is restricted by union or other agreement will be less amenable to accurate rating than will those in which freedom of individual output is unlimited"[1].
3	"Raters will vary in the accuracy of ratings given in direct proportion to the relevancy of their previous contacts with the ratee" [1].	3.a "Close personal friends and relatives of the ratee will be less accurate raters than will close associates on the job only"[1].

4	"Raters will vary in the accuracy of ratings given in direct proportion to the number of previous relevant contacts with the ratee" [1].	4.a "Close job associates will be more accurate raters than will casual acquaintances or infrequent observers"[1]. 4.b "The longer the rater knows the ratee on the job, the greater the probability that the ratings will be accurate" [1]. 4.c "The greater the geographical proximity of the rater to the ratee's workplace, the greater the probability of multiple pertinent experiences and hence, the greater the probability of accurate ratings."[1].

In this construct, Wherry and Barlett explained a dimension of rating which is environmental forces since it affects ratee performance. Wherry and Barlett called for increased freedom of output for the ratee as it allows for accurate evaluations of true ability which means that any interference from external forces should be minimized. Key ideas related to accurate performance evaluation such as providing similar working conditions to all workers, choosing man-paced tasks over machine-paced tasks during evaluations, and minimizing restrictions in rating settings were discussed [1]. They also warned against work unions since they often impose work restrictions.

Wherry and Barlett [1] highlighted the number and relevancy of previous contacts between the ratee and the rater as another rating dimension as it can play a significant role in obtaining accurate evaluation results. According to the ToR [1], relationships between raters and ratees like close friendship should be avoided, close job associates are more accurate raters than will casual acquaintances or infrequent observers. Additionally, the longer the rater knows the ratee on the job, the higher the chances of an accurate rating. To sum up, this construct highlights several dimensions of accurate rating which is providing the ratee with a high degree of controllability, minimizing external restrictions to ratee performance, choosing raters based on the number of relevant contacts between the rater and ratee. Also, raters with a relationship like friendship with a ratee should be avoided, and raters with good amount of knowledge about the ratee's job are preferred to others.

Rating settings

In this construct, several dimensions of rating settings as explained in theorems 8,10, 11, and13, 14, and 15 in addition to hypotheses 11.a, 11.b, and 13.a as shown in table II.

Table II: ToR Theorems and Corollaries

Theorem No.	Definition	Associated Corollaries
8	"If the perceiver is furnished an easily accessible check list of objective cues for the evaluation of performance, to which he can frequently refer, he should be better able to focus his attention properly" [1].	No Associated Corollaries

10	“The keeping of a written record between rating periods of specifically observed critical incidents will improve the objectivity of recall” [1].	No Associated Corollaries
11	“Any setting which facilitates the increase of bias, such as knowledge that the rating will have an immediate effect upon the recipient, will decrease the accuracy of raters, while any setting which stresses the importance to the organization or to society as a whole will decrease perceived bias elements and thus increase accuracy” [1].	11.a "Ratings obtained under experimental conditions (i-e., to be used only to improve instruments, methods, or the like for the good of the organization) will be more accurate than those obtained under actual on-the-job conditions where resulting administrative action will or may affect the ratee." [1]. 11.b “Ratings obtained in advance through a routine process will be more accurate than those especially secured at the time when an administrative action (such as promotion) is contemplated” [1].
13	“Since forgetting is largely a function of intervening activities interposed between learning and recall, ratings secured soon after the observation period will be more accurate than those obtained after a considerable lapse of time” [1].	13.a “A rating should be secured immediately, whenever the ratee’s supervisor is changed in the same job or when the ratee moves to a new position” [1].
14	“If observation is sufficiently frequent so as to constitute overlearning, the accuracy of recall will be improved” [1].	No Associated Corollaries
15	“Observation with intention to remember will facilitate recall” [1]	No Associated Corollaries

Wherry and Barlett [1] postulated that raters can focus their attention during observation if they have a checklist with all the objectives of performance evaluation. Additionally, keeping track of observed critical events in between rating periods can increase the objectivity of recall [1]. Also, they [1] proved that knowledge that the rating is going to affect the ratee can trigger bias. On the other hand, levels of bias decrease when the rater knows that the rating is going to impact the organization, or the society. The ToR [1] proved that accurate ratings should be completed soon after observation to minimize chances of forgetting which results in error in recall. Also, frequent observations can improve the accuracy of recall compared to infrequent ones [1]. To conclude, this construct defines rating settings that yield more accurate ratings, those settings are based on clearly defining evaluation objectives and providing raters with a list of those objectives. Also, to ensure rating settings are free from bias, reasons behind ratings should be related to the organization benefits rather than individual benefits. Additionally, minimizing the time gap between an evaluation report and the observation and conducting frequent observations can foster accurate evaluations.

Rater's bias, knowledge, and trainings

This major construct is covered by theorems 6,7,16, and 17 as well as hypotheses 6.a, 6.b, 7.a, 7.b, 16.a and 17.a as shown by table III.

Table III: ToR theorems and corollaries on rater's bias, knowledge and training

Theorem No.	Definition	Associated Corollaries
6	"The rater will make more accurate ratings when he has been forewarned concerning the types of activity to be rated since this will facilitate their more properly focusing attention on such pertinent behavior" [1].	6.a "Courses for the instruction of raters will be more efficient if they include instruction in what to look for." 6.b "In lieu of such actual instruction, duties which normally involve direct supervisory relation to the ratee, as would be true for an immediate supervisor, will serve to increase rating accuracy" [1].
7	"If the perceiver makes a conscious effort to be objective, after becoming aware of the biasing influence of previous set, he may be able to reduce the influence of his bias" [1].	7.a "Training courses for the rater should include instruction on the effect of set on perception and provide practice in objectivity of observation" [1]. 7.b "Deliberate direction of attention to the objective (measurable) results of behavior may serve to restrain the biasing effects of set" [1].
16	"Performances which are readily classified by the observer into a special category will have relatively larger areal and smaller overall bias components" [1].	16.a "Jobs with simplified performance units requiring a single discrete aptitude will be rated with relatively more areal and less overall bias than will complex jobs requiring a complex pattern of aptitudes" [1].
17	"Rating items which are readily classified by the rater as referring to a given area of behavior will result in relatively larger areal and less overall bias than will items which suggest a complex pattern of behavior to the rater" [1].	17.a "Rating items shown to be factorially unidimensional will result in relatively larger areal and relatively smaller overall bias than will items shown to have a complex factor pattern" [1].

According to Wherry and Barlett [1], bias can lead to under-evaluation or over-evaluation of performance. Evaluation bias was classified into three types: true bias, areal bias, and overall bias. True bias refers to the degree of expectancy of a certain performance or certain quality or ability to be shown by the ratee, this bias is high when the rater has had only a lot of relevant contacts with the ratee [1]. However, areal bias refers to bias that is aroused when the rater sees the ratee in a specific stimulus situation, this situation is categorized by the rater as belonging to a specific behavior, this means areal bias is situation-specific [1]. Overall bias acts as a background bias that does not need a stimulus to get aroused [1]. Wherry and Barlett [1] called for training on ratings and adequate communication of rating objectives with raters as a way to obtain accurate ratings. Moreover, they [1] claimed that acknowledging one's biases and being aware of them can help in endeavors to be objective. The more conscious raters are about the impact of bias on perception, the more deliberate they are in being objective and accurate raters. The ToR also highlights simple and complex behaviors, where simple behaviors are more likely to attract areal bias compared to complex behavior. In sum, this construct explains several types of bias and how to minimize its chances during evaluations. The importance of training and clearly communicating evaluation objectives are highlighted as well as making a conscious effort to be objective.

Rating design

Wherry and Barlett [1] tackled the design of rating instruments in theorems 2,5, 9,12, and 18-27 as well as hypotheses 5.a, 9.a, 9.b, 9.c, 12.a, 12.b, 12.c, and 19.a as shown in table IV below.

Table IV: ToR theorems and corollaries on rating design

Theorem No.	Theorem definition	Associated corollaries
2	"Rating scales or items which have as their behavioral referents those tasks which are maximally controlled by the ratee will lead to more accurate ratings than those which refer to tasks controlled by the work situation" [1].	No associated corollaries
5	"Rating scale items which refer to easily observed behavior categories will result in more accurate ratings than will those which refer to hard-to-observe behavior" [1].	5.a "Rating items which refer to frequently performed acts will be rated more accurately than those which refer to acts performed rarely or at long intervals" [1].
9	"Physical features of a scale which facilitate recall of the actual perception of behavior will increase the accuracy of ratings" [1].	9.a "Longer objective descriptive statements will be more effective than single value words or simple phrases in defining the steps on an adjectival type rating scale" [1]. 9.b "Overall ratings made after completion of a previous objective review (such as would be provided by the previous filling out of a check-list or forced-choice form) will be more accurate than those made without such review" [1]. 9.c "The clearer (more self-explanatory) and more unambiguous the scale to be rated, the more likely that attention will be centered upon the desired behavior" [1].
12	"Knowledge that the rating given will have to be justified may serve unconsciously to affect the rating given" [1].	12.a "Knowledge that the rating may have to be justified to the ratee may cause the rater to recall a higher proportion of favorable perceptions and thus lead to leniency" [1]. 12.b "Knowledge that the rating may have to be justified to the rater's superior may cause the rater to recall a higher proportion of perceptions related

		to actions known to be of particular interest to the superior whether such actions are pertinent or not” [1]. 12.c “To assure that neither of the above distorting effects shall take place alone, it is better to assure their mutual cancellation by requiring that both types of review shall take place” [1].
18	“The effect of adding an increased number of unidimensional items to a single item rating scale is a reduction in random error components, thus giving added relative emphases to true and both areal and overall bias and environmental contamination components” [1].	No associated corollaries
19	“The effect of adding an increased number of items, each from an independent area or factor, to a single item rating scale is a reduction in random error and areal bias components, thus giving added emphasis to true, environmental contamination and overall bias components” [1].	19.a “Of two rating scales, each composed of the same number of items, the one composed of independent items will be more effective than one composed of homogeneous items” [1].
20	“The addition of extra qualified raters, with identical irrelevant contacts with a ratee, on a single item produces the same effect as the addition of extra items, with identical areal classification” [1].	No associated corollaries
21	“The addition of enough extra qualified raters, each with a completely different set of irrelevant contacts with the ratees, will result in the achieving of virtually true ratings in which areal or overall bias as well as error components have disappeared, even though each rater responds but to a single rating item” [1].	No associated corollaries
22	“The effect of adding extra identical items of each type to the items of a heterogeneous scale is to further assure the reduction of error variance but has no increased effect upon the reduction of areal bias. Overall bias is still undiminished” [1].	No associated corollaries
23	The use of several raters on a multi-item homogeneous scale of rating items, when all raters have identical irrelevant contacts with the ratees, has the same effect upon error reduction as multiplying the number of items in the original scale by the number of raters used” [1].	No associated corollaries

24	"The use of several raters on a multi-item completely heterogeneous list of rating items, if all raters' backgrounds are identical, will merely have the same effect as an increase in the number of independent items in the reduction of error variance, but will be least effective in respect to reducing areal bias" [1].	No associated corollaries
25	"To the extent that rater irrelevant contacts with the ratees are somewhat different, the use of plural raters on a completely heterogeneous list of items will result in a reduction of both overall and areal bias variance, with the latter of these two practically disappearing entirely before the former in case the relationship is low" [1].	No associated corollaries
26	"The addition of several extra items to each area of a heterogeneous scale to be used by several raters will further reduce error, but will have no added effect on removal of bias" [1].	No associated corollaries
27	"The use of several raters on a scale composed of several items in each of several areas will further reduce error, but may or may not reduce bias components depending upon the degree of correlation among the irrelevant backgrounds of the raters" [1].	No associated corollaries

According to the ToR [1], rating scales should include rating items that are maximally controlled by the ratee rather than those controlled by the work situation. Since frequently performed actions are easier to observe than rarely performed actions, the physical features of ratings scales can assist in recalling observations and thus increase evaluation accuracy. Wherry and Barlett [1] postulated that ratings scales should deploy long objective and descriptive statements rather than single words or simple phrases. Also, they [1] warned against ambiguous rating scales. The ToR [1] explained the role of justifying ratings in affecting the accuracy of rating results. Consequently, justifying rating scores to the ratee is likely to cause leniency while justifying ratings to the superior of the ratee is likely to facilitate recall of actions related to the superior interests. Finally, they [1] recommended submitting justification to both the ratee and the ratee's supervisor.

Wherry and Barlett [1] demonstrated that rating scales which include multiple instances of a certain behavior may reduce random error. However, these same rating scales could increase the weight of environmental errors and bias components. Thus, they suggested adding unidimensional items that are independent from each other to reduce random error and areal bias but may emphasize environmental errors and other types of bias. Additionally, having different qualified raters with a different set of irrelevant contacts with the ratee can result in more accurate evaluations in which bias is minimum. To sum up, this construct highlights rating design dimensions that help in obtaining accurate ratings. Consequently, it is best to combine independent ratings items and raters with different irrelevant contacts to obtain accurate rating results.

Theory applications in previous literature

The ToR was used in different ways in previous literature, A few researchers used it as part of their conceptual framework [6], and some researchers used it to support overarching claims in their studies [9]. Some studies used the ToR to further explore certain areas that the theory explored [10], [11], and [12]. The ToR was used to study performance appraisal [3], [4], [13], [11], [12], peer evaluation [14], [15] and job interviewers [16].

Performance appraisals (PA) are used to inform administration on personnel decisions and motivate positive work outcomes [11], a method called rater accountability was used to improve the outcomes of PA in [11]. The ToR talks about the same idea of rater accountability and differentiates between two types of accountabilities: upward and downward accountability. Theorem 12 demonstrates the vital role of knowing that ratings should be justified [1] which is the same concept behind rater accountability. The corollaries explain upward and downward accountability and the necessity of combining both to obtain accurate ratings. Many works on PA rely on the ToR when collecting their data [4], [13], they follow corollary 11.a which states that evaluation data collected for research purposes is more accurate than data collected for administrative reasons [1]. Since dyadic relationships can interfere with peer evaluations, many researchers looked into this idea which is an integral part of Wherry’s theory [1]. For instance, some researchers examined how introverts rate their extroverts’ counterparts [13]. Some studies looked at peer evaluation in workplaces and utilized theorem 4, corollaries 4.a and 4.b to select raters [14]. In [15], peer evaluation was studied in the context of different personalities of raters and ratees and the ToR was part of the study’s theoretical framework.

Interviews serve as the most crucial tool for personnel selection, some studies proved that interviewers’ personalities can interfere with their selection decisions of new employees [16], they cited the ToR to support their work. According to [6], teacher observation is prone to different forms of bias: context-dependent bias and context-independent bias, context-independent bias is related to the observer only unlike context-dependent bias which is triggered by the context of evaluation. These types of bias are very similar areal and overall bias in the ToR [1]. Another type of bias called assimilation bias that stems from knowledge about past performance or effectiveness of a ratee was included which is very similar to true bias in the ToR. Table V shows the current application of the ToR and how the theory was utilized, and which theorems or corollaries were used.

Table V Current application of the ToR

Reference	Applications	Use of Theory	Theorems/Corollaries used
[3]	Performance appraisal	Evaluation data collection	Corollary 11.a
[4]	Performance appraisal	Evaluation data collection	Corollary 11.a
[6]	Teacher Observation	Conceptual framework	NA
[13]	Performance appraisal	Evaluation data collection	Corollary 11.a
[11]	Performance appraisal	Exploration for further studies	Theorem 12, corollaries 12.a, 12.b, and 12.c
[12]	Performance appraisal	Exploration for further studies	NA
[14]	Peer Evaluation	Rater Selection	Theorem 4, corollaries 4.a and 4.b
[15]	Peer Evaluation	Conceptual framework	NA
[16]	Job Interviews	Exploration for further studies	NA
[9]	Performance appraisal	Supporting overarching claims	NA
[10]	Peer Evaluation	Exploration for further studies	NA

Implications and future research

This paper explains in detail the first 27 theorems and their hypotheses (corollaries) in the ToR. The remaining theorems require more research and investigation to uncover techniques for testing the reliability of ratings and bias control, most of the techniques recommended by Wherry and Barlett [1] to test the reliability or control bias require a background in statistics and data analysis which is suitable for future work to dive into it. In this section, we seek to highlight the implications of using the ToR in teaching evaluation and highlight key recommendations that can enlighten those in charge; administration personnel who design, set, and interpret evaluations. Teaching evaluation can take many forms, but we focus in this section on evaluation done through a peer instructor or a supervisor, the term evaluation is going to be used synonymously for rating.

Equal evaluation settings among instructors are essential to obtain accurate measurement of the ratee's true ability. Limited resources sometimes create differences in classroom settings and instructors' work environments. However, reducing the gap in work settings is considered one of the pillars for accurate rating [1]. Looking beyond evaluation settings, external forces that can impact ratee's behavior or output pose big threats to accurate evaluation. An example of external forces in teaching context is course coordination. Although course coordination is regarded as an efficient practice and is widely adopted in many schools and universities, course coordination involves rules that are not necessarily set by the instructor which results in restrictions of freedom and may hinder accurate measurement of the instructor's teaching ability.

In terms of an evaluation design, clear objectives and rubrics for evaluation should be provided to supervisors or peers that are going to do an evaluation task for an instructor. Most importantly, selection criteria for evaluating instructors should be based on the number of relevant contacts between the evaluating instructor and the to be evaluated instructor. However, sometimes there is such a limitation that only one instructor teaches a subject at a specific time, this calls for considering instructors who have previous experience with the same subject or a close relation to the field of the subject.

When designing evaluation scales, it is crucial to choose behaviors that are maximally controlled by the instructor and avoid behaviors that may be dictated by external forces like course co-ordination. Also, evaluation scales should include easily observed behaviors like frequently performed behaviors rather than rarely performed ones. Since physical features of an evaluation scale can assist in recalling observed behavior, admin personnel should account for this criterion when designing or choosing evaluation scales. A step towards more accurate observations and evaluation is through asking evaluators to justify their evaluation results. A combination of justifications made to both the evaluated instructor and his/her supervisors can help the evaluator stay focused on the evaluation objectives and minimize error. Moreover, adding multiple unidimensional items to represent a single construct and using independent rating items can contribute to accurate evaluation results. Also, having more than one evaluator simultaneously can reduce bias in the total evaluation result.

Training to enhance objectivity in evaluations should take place regularly. One of the good practices is jotting down notes on critical observed events to assist in recall. Also, evaluators should submit their evaluation responses soon after they are done with the evaluation to minimize chances of forgetting. Minimizing bias can also be achieved through the use of evaluation results in research and development to promote teaching practices and provide feedback to instructors rather than using it to make decisions on awards or punishments like promotion and tenure decisions. The above-mentioned implications provide general suggestions that we believe can help in obtaining accurate teaching evaluation results. We also acknowledge that suggestions should be made with respect to the related contexts.

By looking at the ToR and the current teaching evaluation practice side by side, we can see some similarities and some disparities between them. Table VI shows the standards that appear in ToR as well as current standards of evaluation and the related literature.

Table VI: Current teaching evaluation standards and practices

Standard No.	Teaching Evaluation Standards	Evaluation Standards/ Previous literature
1	Evaluation settings	[17]
2	External forces on performance	[18]
3	Evaluation objectives and criteria	[19], [20]
4	Selection criteria of raters	[21],[20]
5	Evaluation scale design	[17], [18]
6	Justification for evaluation	[17], [19]
7	Use of multiple evaluators	[19]
8	Evaluation uses	[21], [22]
9	Bias training	[24]
10	Evaluation training	[23],[25]

According to the ToR, “tasks in which the raw material, tools, working conditions (light, heat, etc.) are constant from worker to worker will lead to more accurate ratings of ability than will those in which such factors are variable” [1]. Also, current standards of evaluation state that “sound testing practice involves careful monitoring of all aspects of the assessment process and appropriate action when needed to prevent undue disadvantages or advantages for some candidates caused by factors unrelated to the construct being assessed” [17]. The ToR and standards of evaluation are both postulating principles that protect against factors which can create unfair evaluation settings between candidates. Wherry and Barlett advised for performance evaluation of tasks that are controlled by the ratee, they said “tasks in which the performance is maximally controlled by the ratee rather than by the work situation will be more favorable to accurate ratings” [1]. However, recent studies show that teaching evaluations suffer from many shortcomings like measuring factors that are beyond instructors’ control such as student characteristics [18]. Current standards of teaching evaluation need to direct attention to this source of inaccuracy, with collective efforts from researchers, instructors, administrative personnel, more accurate teaching evaluation can be attained.

An important aspect of the ToR that is widely used in teaching evaluation practice is the use of evaluation rubrics. Wherry and Barlett said “if the perceiver is furnished an easily accessible check list of objective cues for the evaluation of performance, to which [they] can frequently refer, [they] should be better able to focus [their] attention properly” [1]. The use of evaluation rubric in class observations to evaluate teaching is a well-established practice in previous literature [19], [20]. The ToR explains the role of choosing the appropriate raters to evaluate a ratee, they said “Raters will vary in the accuracy of ratings given in direct proportion to the number of previous relevant contacts with the ratee”. Several peer teaching evaluation studies reported peer matching based on the taught subject [21], years of experience [20]. Such an approach of peer matching in teaching evaluation context is driven by data that can support relevancy between the rater and ratee as stated by the ToR.

Moving to evaluation scale design, whether evaluation scales are developed by the institution or purchased from an external evaluation entity, with evidence on teaching evaluation measuring things beyond instructor control [18], there are more concerns about evaluation scales, their use and interpretation from the administration side. According to the standards of evaluation, experience and sound level of experience should be used to interpret results [17]. Since we know from the ToR that evaluation scales need to consider behaviors that are controlled by the ratee [1], more evidence is needed to understand how to realize this concept in teaching evaluation context with so many factors such as student characteristics, teacher experience, grade taught, etc.

In current evaluation standards, supporting the validity of evaluation requires [17]. Similarly, accurate evaluation is supported by justification as well as the assignment of multiple evaluators according to the ToR [1]. Teaching evaluation studies reported the use of multiple sources of evidence as a way to support teaching evaluation decisions [19]. In [17], evaluation uses are determined by organizational values. However, the ToR explains thoroughly the effect of evaluation use on the accuracy and objectivity of its results, Wherry and Barlett said that "ratings obtained under experimental conditions (to be used only to improve instruments, methods, or the like for the good of the organization) will be more accurate than those obtained under actual on-the-job conditions where resulting administrative action will or may affect the ratee". Several studies on teaching evaluation criticize the use of teaching evaluations in making promotion and tenure decisions, especially student evaluations of teaching [22]. While many instructors have some trust in peer evaluations of teaching, many of them do not believe in student evaluation of teaching [21]. Careful attention to the uses of teaching evaluation is needed if the accuracy of teaching evaluations is sought.

To maintain accurate evaluation results, the ToR calls for training on objectivity to increase awareness on the effect of perception on the accuracy of evaluation results, Wherry and Barlett said "training courses for the rater should include […] practice in objectivity of observation" [1]. Recent studies include efforts to reduce bias in peer evaluations of teaching [22] and student evaluations of teaching [24]. In general, researchers are realizing a need to focus attention to training on teaching evaluation, they are also interested to understand how evaluation training can contribute to teaching evaluation success [23],[25]. According to Wherry and Barlett, training on how to evaluate accurately is one of the big constructs of the ToR [1].

To sum up, some of the ToR principals are well established in teaching evaluation standards like the use of evaluation rubrics while some principals are currently drawing researcher's attention like training on teaching evaluation and bias training. Also, some principals are not yet realized in teaching evaluation standards such as eliminating external forces that are beyond instructor's control. In this comparison, we sought to draw our reader's attention to current practice of teaching evaluation with respect to the ToR to highlight the theory's meaning and contribution to teaching evaluations as well as highlight areas in which the theory can be leveraged to conduct more research and achieve more improvement.

Conclusion

The ToR offers a comprehensive guide to designing and interpreting performance evaluations, it has many applications in professional and non-educational fields with the objective of increasing organizational outcomes such as work efficiency. Although some of the theory's principals appear in current teaching evaluation standards, some principals are still lacking. That is why we believe that the ToR needs more attention from educational scholars since it covers many dimensions of evaluation systems like rating design, rating settings, and rating conditions. The ToR talks in detail about many aspects of the evaluation scene like the relationship between the rater and ratee, the degree of ratee's control, ratings settings, ratings design, bias training, testing reliability of evaluation and bias control. This paper aimed at explaining four out of six major theory constructs which span 27 theorems and their corollaries. Also, applications for the theory in previous literature and its potential enlightenment in reforming teaching evaluation were explored too. Finally, a comparison was drawn between current teaching evaluation standards and the ToR.

Appendix I

The whole body of theory was formulated by Wherry and Barlett (1952) quantitatively with lots of variables, weights, and constants. To increase the readability of this report, only one equation will be shown to draw some light on the original sense of epistemology and methodology in this theory. The ToR used the mental test theory to describe ratee performance as dependent on three things: ratee ability, environmental factors, and random error. So, the ratee performance equation can be described using Eq.1.

$Z_{X_A} = t_A . Z_T + i_A Z_I + e_A . Z_{EA} \rightarrow$ (Eq.1)
Where Z_{X_A} is the performance of ratee, Z_T is the true ability, Z_I are the environmental factors. And Z_{EA} is random error, whereas, t_A,i_A, and e_A are the weights of the different components.

References

[1] R. J. Wherry and C. J. Bartlett, "THE CONTROL OF BIAS IN RATINGS: A THEORY OF RATING," Personnel Psychology, vol. 35, no. 3, pp. 521–551, Sep. 1982, doi: 10.1111/j.1744-6570.1982.tb02208.x.

[2] Y. Chuang, H. Chiang, and A. Lin, "Helping behaviors convert negative affect into job satisfaction and creative performance: The moderating role of work competence," PR, vol. 48, no. 6, pp. 1530–1547, Sep. 2019, doi: 10.1108/PR-01-2018-0038.

[3] M. Kilduff, A. Mehra, D. A. (Denny) Gioia, and S. Borgatti, “Brokering Trust to Enhance Leadership: A Self-Monitoring Approach to Leadership Emergence,” in Knowledge and Networks, vol. 11, J. Glückler, E. Lazega, and I. Hammer, Eds., in Knowledge and Space, vol. 11. , Cham: Springer International Publishing, 2017, pp. 221–240. doi: 10.1007/978-3-319-45023-0_11.

[4] A. Speer, “Quantifying with words An investigation of the validity of narrative-derived performance scores.” Personnel psychology, 2018.

[5] X. M. Wang, K. F. E. Wong, and J. Y. Y. Kwong, “The roles of rater goals and ratee performance levels in the distortion of performance ratings.,” Journal of Applied Psychology, vol. 95, no. 3, pp. 546–561, 2010, doi: 10.1037/a0018866.

[6] S. B. Hunter, “The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores,” AERA Open, vol. 6, no. 2, p. 233285842092927, Apr. 2020, doi: 10.1177/2332858420929276.

[7] M. Moeller, “Running head: PEER GRADING”.

[8] R. J. Wherry, “The control of bias in rating: A theory of rating,” Washington, DC: Department of the Army, Adjutant General’s Office, Personnel Research Section, 1952.

[9] S. H. Moon, S. E. Scullen, and G. P. Latham, “Precarious curve ahead: The effects of forced distribution rating systems on job performance,” Human Resource Management Review, vol. 26, no. 2, pp. 166–179, Jun. 2016, doi: 10.1016/j.hrmr.2015.12.002.

[10] J. A. Schmidt, T. A. O’Neill, and P. D. Dunlop, “The Effects of Team Context on Peer Ratings of Task and Citizenship Performance,” J Bus Psychol, vol. 36, no. 4, pp. 573–588, Aug. 2021, doi: 10.1007/s10869-020-09701-8.

[11] A. P. Tenbrink and A. B. Speer, “Accountability during Performance Appraisals: The Development and Validation of the Rater Accountability Scale,” Human Performance, vol. 36, no. 1, pp. 1–23, Jan. 2023, doi: 10.1080/08959285.2021.2023876.

[12] T. Vriend, C. Rook, H. Garretsen, J. I. Stoker, and M. Kets De Vries, “Relating Cultural Distance to Self-Other Agreement of Leader–Observer Dyads: The Role of Hierarchical Position,” Front. Psychol., vol. 12, p. 738120, Oct. 2021, doi: 10.3389/fpsyg.2021.738120.

[13] R. Aalbers, “J of Product Innov Manag - 2015 - Aalbers - Vertical and Horizontal Cross-Ties Benefits of Cross-Hierarchy and Cross-Unit.pdf.” Journal of Product Innovation Management, 2015.

[14] O. L. Clark, M. J. Zickar, and S. M. Jex, “Role Definition as a Moderator of the Relationship Between Safety Climate and Organizational Citizenship Behavior Among Hospital Nurses,” J Bus Psychol, vol. 29, no. 1, pp. 101–110, Mar. 2014, doi: 10.1007/s10869-013-9302-0.

[15] A. Erez, P. Schilpzand, K. Leavitt, A. H. Woolum, and T. A. Judge, “Inherently Relational: Interactions between Peers’ and Individuals’ Personalities Impact Reward Giving and Appraisal of Individual Performance,” AMJ, vol. 58, no. 6, pp. 1761–1784, Dec. 2015, doi: 10.5465/amj.2011.0214.

[16] W. Tsai, F. HsinHung Chen, H. Chen, and K. Tseng, “When Will Interviewers Be Willing to Use High-structured Job Interviews? The role of personality,” Int J Selection Assessment, vol. 24, no. 1, pp. 92–105, Mar. 2016, doi: 10.1111/ijsa.12133.

[17] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 2014.

[18] S. L. Campbell and M. Ronfeldt, “Observational Evaluation of Teachers: Measuring More Than We Bargained for?,” American Educational Research Journal, vol. 55, no. 6, pp. 1233–1267, Dec. 2018, doi: 10.3102/0002831218776216.

[19] K. A. Villanueva, S. A. Brown, N. P. Pitterson, D. S. Hurwitz, and A. Sitomer, “Teaching Evaluation Practices in Engineering Programs: Current Approaches and Usefulness,” 2017.

[20] J. Papay, “Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data”.

[21] L. Cutroni and A. Paladino, “Peer-ing in: A systematic review and framework of peer review of teaching in higher education,” Teaching and Teacher Education, vol. 133, p. 104302, Oct. 2023, doi: 10.1016/j.tate.2023.104302.

[22] C. Ferekides et al., “Breaking Boundaries: An Organized Revolution for the Professional Formation of Electrical Engineers,” in 2022 ASEE Annual Conference & Exposition Proceedings, Minneapolis, MN: ASEE Conferences, Aug. 2022, p. 41972. doi: 10.18260/1-2--41972.

[23] C. J. Arévalo Gross, P. Rodríguez-Bilella, and C. Olavarría, “How to Train Better in Evaluation: Teaching Landscape and Lessons Learned From Latin America,” American Journal of Evaluation, vol. 44, no. 2, pp. 282–292, Jun. 2023, doi: 10.1177/10982140211059373.

[24] D. A. M. Peterson, L. A. Biederman, D. Andersen, T. M. Ditonto, and K. Roe, “Mitigating gender bias in student evaluations of teaching,” PLoS ONE, vol. 14, no. 5, p. e0216241, May 2019, doi: 10.1371/journal.pone.0216241.

[25] J. M. LaChenaye, A. S. Boyce, J. Van Draanen, and K. Everett, “Community, Theory, and Guidance: Benefits and Lessons Learned in Evaluation Peer Mentoring,” Canadian Journal of Program Evaluation, vol. 34, no. 1, pp. 102–117, Mar. 2019, doi: 10.3138/cjpe.42118.