Impact of educating faculty on student assessment: beyond satisfaction level

Wessam Mohamed
Centre for Quality of Assessment and Examinations,
Vice Presidency Office for Academic Affairs,
Imam Abdulrahman Bin Faisal University,
Dammam, Saudi Arabia

Abstract

Purpose – This study evaluated the impact of a faculty training program on student assessment using the Kirkpatrick model.

Design/methodology/approach – A self-reported survey assessed 111 Saudi and non-Saudi participants’ satisfaction. Subjective and objective measures (self-reported measures, assessment literacy inventory and performance-based assessment tasks) gauged participants’ learning level. Pre- and post-training data were collected from 2020 to 2022.

Findings – A highly significant effect on satisfaction (>80%) and learning levels was observed, as manifested by workplace practices of student assessment (>70%, the cut-off score). Pre- and post-training comparisons of participants’ satisfaction and assessment literacy scores showed significant improvements following training. Multiple regression analyses showed no significant effects for gender and educational attainment but a substantial impact of academic cluster on participants’ student assessment skills.

Research limitations/implications – Long-term effects of training faculty on assessment practices and student achievement will be studied at the institutional level in future research.

Practical implications – The current study contributes to human capital investment via faculty training on student assessment, helping them comply with assessment best practices. This assures the quality, fairness and consistency of assessment processes across disciplines in higher education institutions, enhances assessment validity and trust in educational services and may support institutional accreditation.

Social implications – This study provides opportunities for sharing best practices and helps establish a community of practice. It enhances learning outcomes achievement and empowers higher education graduates with attributes necessary to succeed in the labor market. The human capital investment may have a long-term impact on overall higher education quality.

Originality/value – This study contributes to the scarce literature investigating the impact of training faculty from different clusters on student assessment using subjective and objective measures. It provides developing and evaluating a long-term student assessment program following the Kirkpatrick model.

Keywords Assessment literacy, Evaluation, Kirkpatrick model, Professional development

Paper type Research paper

The completion of this research could not have been possible without the substantial contribution and assistance of many persons and entities at the university. The author would like to express deep gratitude to the Vice Presidency of Academic Affairs and the Director of the QA center, whose visionary leadership, commitment to excellence and seamless support have been a source of inspiration throughout the journey of the CSA program since its launch. The author’s sincerest acknowledgment extends to the Heads of Units at the QA Centre, who have played a pivotal role in the realization of the CSA program since 2018. Their endeavors, under the supervision of the QA director (also the CSA executive committee), in shaping and establishing the CSA have been instrumental. The author wants to express wholehearted appreciation to the trainers and assessors whose dedication and expertise ensured the integrity and rigor of the CSA implementation. Finally, the author is profoundly thankful to the colleagues who participated in the CSA program and generously allowed the author to use their data and responses for research and publication. This acknowledgment reflects the collective efforts of all the above-mentioned, whose assistance has greatly enriched the CSA experience.
1. Introduction
Higher education institutions (HEIs) are committed to delivering high-quality education, including student assessment (SA), to meet labor market demands (Boyle and Bowden, 1997). HEIs seek to promote capable faculty through professional development (PD) that offers sustained formal learning opportunities (e.g. Daumiller et al., 2021). PD is defined as “structured professional learning that results in changes to teacher knowledge, practices . . .” (Darling-Hammond et al., 2017, p. 2). PD offers ongoing educational opportunities for faculty, who are expected to continually update their knowledge and skills to remain competent and effective (Guskey and Yoon, 2009). One of educators’ roles is conducting SA to gauge student success, but many educators lack the requisite knowledge and skills (DeLuca, 2012; Jawhar and Subahi, 2020). Recognizing the pivotal link between teacher quality and student success (Shah et al., 2023), HEIs view SA as a potential training program in PD, guided by the Compass Model (Al-Eraky and McLean, 2012). Although researchers have explored multiple approaches to assess the impact of faculty development (FD) programs (Senel, 2014), whether they have short (ideally less than one day) or long duration (Daumiller et al., 2021), trainee satisfaction remains the most common indicator of program success (Muammar and Alkathiri, 2021). Among the models used to assess FD training effectiveness in higher education, Kirkpatrick and Kirkpatrick’s (2006) model is the most influential. It assesses FD programs on four levels: reaction/satisfaction, learning, behavioral changes, and results (Kirkpatrick and Kirkpatrick, 2006).

Despite its limitations (Cahapay, 2021), the model has several advantages, including its widespread use and adaptability across diverse disciplines (Alsalamah and Callinan, 2021). The model can contextualize findings because the four outcome levels require descriptive or evaluative data collection (Cahapay, 2021), and its success in monitoring student learning outcomes and assessing educational activities (Alsalamah and Callinan, 2021; Shah et al., 2023) has been demonstrated, further emphasizing its utility.

However, the Kirkpatrick model has been insufficiently explored in evaluating SA-related FD programs, particularly in the Saudi higher education context. Therefore, the current study aims to fill this gap by assessing the effectiveness of educating faculty through the “Certificate of Student Assessment” (CSA) training program using the Kirkpatrick model, specifically targeting assessment literacy (AL). AL, imperative for teachers’ development (Gotch and French, 2014), positively influences their teaching and learning (T&L) practices (Ashraf and Zolfaghari, 2018) and connects assessment quality and student achievement (Mellati and Khademi, 2018). Furthermore, this study recognizes the inadequacy of trainee satisfaction as the sole measure used to examine the impact of educating faculty via PD programs. Therefore, it adopts the Kirkpatrick model to determine the impact of faculty education through a long-term FD program (CSA) by assessing faculty satisfaction and learning levels. Additionally, the study examines the factors contributing to faculty’s SA-related skills by addressing three questions:

*RQ1.* What are the satisfaction levels of faculty after they received SA training, as assessed by subjective measures?

*RQ2.* What are the learning levels of faculty after receiving SA training, as assessed by subjective and objective measures?

*RQ3.* Which demographic variables (gender, academic cluster, and educational qualification) affect faculty learning levels, as measured by the assessment literacy inventory (ALI)?

The research outcomes contribute to achieving CSA objectives, such as helping faculty improve assessment practices in the T&L process and performance at work. Notably, the terms “educational assessment skills” and “assessment literacy” are used synonymously in this paper; “educating faculty,” “training,” “faculty development,” and “professional development” are used interchangeably.
2. Literature review

2.1 Educational assessment of students

Assessment is critical in the T&L process, as teachers spend up to half of their professional time on assessment (Stiggins, 2018). Assessment is defined as the process of gathering and interpreting students’ responses to an educational task for various purposes, including assigning grades (Brink and Lautenbach, 2011; Harlen et al., 1992) or shaping students’ understanding of the curriculum and supporting further progress (Abduh, 2021; Gikandi et al., 2011). Teachers can observe the impact and effectiveness of their instructional approaches and strategies through assessments (Jawhar and Subahi, 2020). Plake et al. (1993) voiced their concerns about the adequacy of teachers’ assessment skills more than 30 years ago. Research shows that many teachers perform SA practices without possessing the expected knowledge (e.g. DeLuca, 2012). In response, three associations – the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association – developed the Standards for Teacher Competence in Educational Assessment of Students (American Federation of Teachers and Association, 1990; Brookhart, 1999; Stiggins, 1995). These standards have influenced T&L practices and FD programs (Brookhart, 2011), and are intended to serve as a guide for student assessors and a self-assessment tool for teachers (American Federation of Teachers and Association, 1990). Additionally, specific behaviors reflecting AL were articulated in seven major competency areas of educational assessment skills (American Federation of Teachers and Association, 1990; Brookhart, 2011). DeLuca et al. (2016) emphasized that the Standards remain the predominant framework for creating AL and its instruments. Recently, scholars have stressed that teachers must use AL to effectively implement classroom-based assessment practices, make accurate and informed decisions regarding their students’ achievements, and contribute to educational reform (Gotch and French, 2014; Jawhar and Subahi, 2020). To measure teachers’ competencies in assessing students, Mertler and Campbell (2005) developed the ALI and evaluated its psychometric properties, which has been used to evaluate teachers’ preparedness to administer SA by demonstrating an appropriate level of skills (Iqbal et al., 2023; Jawhar and Subahi, 2020; Levy-Vered and Nasser-Abu Alhija, 2015; Saleem et al., 2022).

Substantial research has evaluated the AL of teachers in pre-service programs and educational institutions; it has also examined the impact of discrete courses on their AL (DeLuca, 2012) and its application in workplace contexts. The results highlight inadequate overall AL levels and poor teacher performance (e.g. Afshar and Ranjbar, 2021; Najib Muhammad and Bardakçı, 2019), concluding that junior teacher candidates are not sufficiently prepared to assess student learning, and that assessment remains a significant gap in pre-service programs. Further, some studies assessing AL at the college level have found that teachers in Saudi and Pakistani contexts report inadequate AL (Iqbal et al., 2023; Jawhar and Subahi, 2020; Saleem et al., 2022).

Although several studies show that AL levels are related to specific variables, such as teaching experience (Mertler, 2003) and academic discipline (Jawhar and Subahi, 2020), others report no significant effects related to gender (Jawhar and Subahi, 2020; Saleem et al., 2022), educational qualifications, or years of experience (Jawhar and Subahi, 2020), presenting inconsistent results. Additionally, researchers have recommended the development of related training in this area to enhance AL in different disciplines in the higher education context (e.g. Saleem et al., 2022). Battistone et al. (2019) concluded that teacher education programs and educational institutions must be accountable for scaffolding early career teachers via FD in competencies such as constructively aligning curriculum, instruction, and assessment. Increased performance (Mertler, 2003) and confidence in assessment practices (DeLuca and Bellara, 2013) were observed following education, training, or experience with SA.

Given the relevance of ensuring adequate SA skills via FD programs and the paucity of relevant research, especially in new contexts (Gotch and French, 2014), this study’s contribution is to evaluate the impact of SA-based FD programs in Saudi higher education.
2.2 Evaluation of the impact of PD

Traditionally, PD is defined as a broad range of activities or planned programs that institutions develop or use to support faculty in performing their duties and to improve their knowledge and skills (Steinert, 2014). In this regard, PD is considered a significant determinant in teacher quality (Muammar and Alkathiri, 2021; Steinert, 2014), which has been empirically linked to students’ academic success (e.g. Batezai and Kiazai, 2022). Instructors with a higher level of education, experience, knowledge, and skills have more significant and positive impacts on students’ academic achievements (e.g. Nazar and Nordin, 2020; Shah et al., 2023). PD programs have been conducted to improve teacher quality through T&L, curriculum development, e-learning, research, technological enhancement, administration, leadership development (Muammar and Alkathiri, 2021; Steinert, 2014), and quality of SA practices in HEIs (Bibler Zaidi et al., 2017; Elassy, 2015; Hassanien, 2018). Existing research has shown that PD has a measurable impact on instructional practices (Condon et al., 2016), as well as improvements in SA practices (Bibler Zaidi et al., 2017), instructional practices, and student learning (e.g. Condon et al., 2016), particularly among teachers who participated in long-term PD activities (Daumiller et al., 2021). Moreover, providing PD opportunities to faculty often helps HEIs receive programmatic accreditation (Alkathiri, 2020).

To evaluate the impact of PD programs systematically, participant satisfaction measures are commonly utilized (Muammar and Alkathiri, 2021). However, other approaches to evaluating continuing education programs by demonstrating their effectiveness have been proposed. Kirkpatrick and Kirkpatrick (2006) formulated an outcome-based four-stage evaluation model in 1959, which is now widely used. Some scholars have criticized the model (Cahapay, 2021); however, Alsalamah and Callinan’s (2021) bibliometric analysis of 416 articles on Kirkpatrick’s model published between 1959 and July 2020 revealed that the model is adequate, applicable, and adaptable in various FD contexts. Malik and Asghar (2020) called the model a good choice for assessing education and training program effectiveness immediately after program introduction because of the four levels of outcomes. In the Kirkpatrick model, Level 1 (reaction/satisfaction) refers to participants’ satisfaction and reception of the program and its components. Level 2 (learning) shows how participants’ knowledge, attitudes, or skills improve because of program participation. Level 3 (application of learning) shows behavioral changes in the professional context or workplace. Finally, Level 4 (results) identifies the training’s impact on the organization, program, and outcomes concerning the goals and objectives (Kirkpatrick and Kirkpatrick, 2006). Johnston et al. (2018) analyzed studies evaluating simulation and debriefing in health care using the Kirkpatrick model. They found that most studies measured either Level 1 or 2 using self-reported scales, skills rubrics, score improvement from pre-test to post-test, or gains over time, as observed in the intervention group. However, studies showed either no significant transfer of learning to real practice (Level 3) or the absence of research demonstrating the effect of learners’ actions on outcomes (Level 4) due to factors such as the long-term nature of research evaluating outcome changes (Johnston et al., 2018). Ghasemi et al. (2020) evaluated the effectiveness of a scientific writing workshop using the Kirkpatrick model and showed participants’ high levels of satisfaction with the workshop (Level 1), improvements before and after the workshop (Level 2), and knowledge transfer that translated into research projects and publications (Levels 3 and 4). Importantly, the authors did not find a gender effect in the improvement of participants’ knowledge. Extending the use of the Kirkpatrick model to other disciplines, Mahmoodi et al. (2019) evaluated in-service PD programs for English language teachers and showed a positive effect. Recently, the Kirkpatrick model was used to evaluate the effectiveness of an online course and achievement of its learning outcomes (Alsalamah and Callinan, 2021). The progress of 305 Pakistani teachers following their participation in an accelerated learning program (Shah et al., 2023) was also examined using the model, resulting in positive responses to the program (Level 1) and progress in teaching methods, communication skills, and academic performance (Level 4).
Studies show the Kirkpatrick model as a valuable tool for assessing educational efforts, FD program performance, and the impact of educational activities in various contexts (Shah et al., 2023). However, the planning and evaluation of long-term FD programs in SA using the Kirkpatrick model has rarely been studied. Therefore, the present study examines the effects of a newly designed FD program in SA using subjective and objective measures to determine participant satisfaction and learning levels. Thus, the following hypotheses were proposed:

$H1$. Educating faculty on SA enhances their satisfaction levels, measured subjectively.

$H2$. Educating faculty on SA enhances their learning, measured subjectively and objectively.

$H2a$. Educating faculty on SA enhances their learning levels, measured subjectively.

$H2b$. Educating faculty on SA enhances their learning levels, measured objectively.

$H3$. Gender, academic cluster, and educational qualification significantly affect post-training learning levels, as measured by the ALI.

3. Methodology
This study evaluated the impact of training educators via a long-term FD program (CSA) constructed according to the concept of AL. Only Levels 1 and 2 of the Kirkpatrick model were measured; Levels 3 and 4 will be assessed in a future study to be conducted in the next 1–2 years.

3.1 Design of SA-based FD program for educating faculty
CSA was developed by members of the Quality of Assessment and Examinations Center (QA Center), an institution affiliated with the Vice Presidency for Academic Affairs (VPAA) of a Saudi college. It underwent a lengthy revision process from 2018 until its initial release in 2020. Following Kirkpatrick and Kirkpatrick (2006), ten factors were carefully considered in creating and implementing CSA as follows:

1. Determining needs: Data were analyzed from two resources: institutional and strategic needs, training needs assessment surveys and performance indicator reports of universities and colleges.

2. Setting objectives: The CSA program was motivated by the VPAA’s strategic goals of enhancing the caliber of academic programs through qualified faculty members who attend FD programs; its primary purpose is educating faculty on SA.

3. Determining subject content: CSA comprises two modules, totaling 50 training hours (Figure 1), and includes assessment tasks. Its content aligns with that of AL (Brookhart, 2011; Mertler and Campbell, 2005).

4. Selecting participants: The heads of QA units and committees at the colleges were asked to become CSA recipients and knowledge-transfer agents. Participants received training based on their academic clusters.

5. Determining the best schedule: CSA was offered as a long-term program (one week per module) spread over two months. The program was announced approximately one month prior to its implementation with a one-day orientation and pre-assessment.

6. Selecting appropriate facilities: The training environment was a well-equipped and comfortable space that facilitated group work.
(7) Selecting and preparing audio-visual aids: CSA was conducted face-to-face (F2F), and trainees were assessed either F2F or online through the Blackboard learning management system (BB).

(8) Selecting appropriate instructors: Qualified instructors with experience in training, teaching, assessment, and FD program development were selected. Training methods, including interactive lectures, discussions, group work, case studies, hands-on exercises, and flipped classrooms, promoted the principles of self-directed cooperative learning and motivation. Formative assessment was also used to ensure active participation.

(9) Coordinating CSA: QA administrative teams (employees) were assigned to coordinate logistics.

(10) Evaluating the Program: The QA Center assessed the CSA’s impact using the Kirkpatrick model.

**Figure 1.** Structure of CSA program

**Note(s):** The image was taken from “An Introductory Handbook to Certificate of Student Assessment Program (CSA),” Second Edition, 2021-2022 (p.8), by the Centre for the Quality of Assessment and Examinations
3.2 Sample
Purposive sampling was employed to select 111 participants aged 30–60 years as a homogeneous group with a specific college position at a medium-sized Saudi university in the Eastern Province. Participants were heads of either units or committees for quality assessments and examinations who fulfilled the requirements of CSA completion. The sample included Saudi (45%) and non-Saudi (55%) individuals, which reflects the context of Saudi HEIs. The same institutional assessment practices were applied to both categories. Detailed demographic information (academic cluster, gender, and educational qualification) was collected using a self-report questionnaire (Table 1). Data collection was voluntary and confidential. The participants who answered yes to the question “Would you agree to the use of your responses in the CSA program for research and publishing purposes?” were included in the study. The participation rate was 99.1%.

3.3 Instruments
3.3.1 Satisfaction survey. Since 2018, the QA Center has conducted a validated self-report survey to measure response/satisfaction with six key components of training programs based on the Kirkpatrick model (Kirkpatrick and Kirkpatrick, 2006). The survey includes six items (Table 2) that address the major constructs reported as having the greatest influence on

| Table 1. Frequencies and percentages of sample characteristics ($n = 111$) |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| Academic cluster     | Sub-group            | Frequency | Percent (%) |
| Health               | 32                   | 28.8%     |
| Engineering, science, and administration | 39 | 35.2% |
| Arts and humanities  | 40                   | 36%       |
| Gender               | Men                  | 51        | 45.9%     |
|                       | Women                | 60        | 54.1%     |
| Educational qualification | Obtained Ph.D     | 21        | 18.9%     |
|                       | MA or BA             | 90        | 81.9%     |
| Source(s): The author’s own work |

| Table 2. Descriptive statistics of self-reported satisfaction and learning levels (number of responses = 711) |
|---------------------------------------------------------------|----------------------|----------------------|----------------------|----------------------|
| Item                                                          | 1 | 2 | 3 | 4 | 5 | M | SD | Skewness (Pearson, Fischer) |
| 1-Training was well organized                                 | 0.3 | 1 | 3 | 21.7 | 74 | 4.68 | 0.61 | -0.53, -25.03 |
| 2-Workshops helped me acquire the skills needed for SA        | 0.6 | 0.6 | 5.5 | 24.2 | 69.1 | 4.61 | 0.67 | -0.58, -25.03 |
| 3-Content was appropriate to my training needs                | 0.4 | 0.8 | 6.7 | 20.5 | 71.6 | 4.62 | 0.69 | -0.56, -21.82 |
| 4-Trainer was qualified enough to deliver the workshop        | 0.6 | 1.1 | 3.3 | 17.4 | 77.6 | 4.70 | 0.64 | -0.60, -29.71 |
| 5-Facilities/training rooms were satisfactory                 | 0.4 | 0.4 | 3.2 | 22.2 | 73.8 | 4.69 | 0.59 | -0.47, -25.46 |
| 6-Workshop is applicable to real job situations               | 0.6 | 0.6 | 3.5 | 22.9 | 72.4 | 4.66 | 0.63 | -0.54, -25.46 |
| Overall satisfaction                                          | 0.4 | 0.7 | 4.2 | 21 | 73.7 | 4.67 | 0.63 | -0.52, -24.77 |
| 7-I would rate the level of my prior skills of this topic as...| 3.8 | 6.9 | 25.3 | 30.5 | 33.5 | 3.83 | 1.09 | 0.16, 7.60 |
| 8-I would rate the level of my skills of this topic after this training as... | 0.7 | 1.3 | 4.9 | 25.2 | 67.9 | 4.58 | 0.71 | -0.59, -22.63 |
| Note(s): *M* mean, *SD*: standard deviation                   |
| Source(s): The author’s own work                              |
participants’ positive evaluation of training programs (Muammar and Alkathiri, 2021). The survey was translated from English to Arabic by the author and back-translated by a professional translator. Ten experts evaluated the items’ content validity, clarity of wording, and equivalence of Arabic and English phrasing. The experts were selected based on their expertise in SA, medical education, teaching experience in delivering SA-related workshops and classes, research and publication records in SA, and experience in developing scales in SA-related areas. Based on their feedback, minor changes were made. Participants were asked to rate their satisfaction after each training day on a 5-point Likert scale (5 = strongly satisfied to 1 = strongly dissatisfied) A pilot study was conducted with three colleagues to check their understanding of the survey items and ensure the translation’s accuracy and validation. Confirmatory factor analysis (CFA) with the maximum likelihood estimation method was used in Amos (Version 26.0) Arbuckle, 2019) to assess the construct validity of the latent construct (Zainudin, 2015). The CFA indices for the one-factor model were good (CMIN/DF = 4.6, RMSEA = 0.07, SRMR = 0.017; GFI = 0.99, AGFI = 0.96, CFI = 0.99, NFI = 0.99; IFI = 0.99, TLI = 0.98, AIC [default model = 60.33, less than the independence model = 2932.76]) (Byrne, 1998; Hu and Bentler, 1999; Zainudin, 2015). The scale was observed to have good reliability (Cronbach’s α = 0.91; inter-item correlation mean = 0.63), while the composite reliability (CR) and average variance extracted (AVE) were acceptable at 0.89 and 0.53, respectively (Abdelrahman et al., 2022). The questionnaire was created using “Question-Pro” and distributed online.

3.3.2 Learning level measures. 3.3.2.1 ALI. ALI was originally developed by Mertler and Campbell (2005) to parallel the Standards for Teacher Competence in the Educational Assessment of Students through five scenarios. Each scenario was followed by seven multiple-choice questions (MCQs) with four options, of which only one was correct. The seven assessed standards were Choosing Appropriate Assessment Methods (CAAM); Development of Appropriate Assessment Methods (DAAM); Administration, Scoring, and Results Interpretation (ASRI); Using Results for Making Decisions (URMD); Using Results of Assessment for Grading (URAG), Communication of Assessment Results (CAR); and Recognizing Universal Methods of Assessment (RUMA). ALI is still commonly used in different contexts and has been reported to have acceptable psychometric properties for content validation and reliability (Akhtar et al., 2021; Iqbal et al., 2023; Jawhar and Subahi, 2020; Levy-Vered and Nasser-Abu Alhija, 2015; Mertler and Campbell, 2005; Saleem et al., 2022). The translation and adaptation included four main steps, following Levy-Vered and Nasser-Abu Alhija (2015): (1) forward translation [1] into Arabic; (2) expert panel back-translation to ensure accuracy and internal consistency; (3) content validity of the translated version by six faculty members specializing in instruction, educational measurement, and medical education, who also work as professional trainers in SA. They judged the clarity of the wording, appropriateness, and relatedness of the scenarios and items against the assessed construct (standard). The survey was refined based on agreement and feedback (Alkharusi, 2011a) and shortened to three scenarios with 21 items; (4) a pilot test of the shortened version was presented to faculty with different AL levels. Moreover, the author examined the model fit of the seven-factor model of ALI. The CFA model goodness-of-fit was examined using three indices, showing that the model is acceptable (CMIN/DF = 1.21, RMSEA = 0.045, SRMR = 0.0169, GFI = 0.86, AGFI = 0.81, CFI = 0.99, NFI = 0.99; PGFI = 0.64; AIC [default model = 326.69, less than the independence model = 330.15]) (Byrne, 1998; Hu and Bentler, 1999; Zainudin, 2015). Finally, the item analysis for MCQs was conducted to ensure the items met the minimum quality control criteria (Varma, 2006) according to the difficulty and discrimination indices (Alkharusi, 2011b; Crocker and Algina, 1986; Varma, 2006). Overall, the items’ difficulty indices ranged from 10.3% to 96.3% (average of 53%), indicating that items are in the appropriate range for the measured construct (Alkharusi, 2011b). Analyses of item discrimination indicated that the items differentiated between high and low performers.
None of the items were reported to have negative discrimination, with the values ranging from 0.03 to 0.48 (average of 0.28). Additionally, the item-total correlation coefficients for approximately 86% of the items were statistically significant (0.036–0.37), suggesting that most items contributed positively and significantly to the overall score; the reliability coefficient using Cronbach's alpha was 0.4. Although the value is low, it is consistent with a previously reported low-reliability value (0.54) generated for the 35-item ALI for a sample of in-service teachers (Plake et al., 1993). The final version was delivered in a computerized format via BB. The examination duration was one hour, allowing only one attempt. The items were scored dichotomously (1 = correct response, 0 = incorrect response, maximum score: 21 [100%]), with higher total scores reflecting higher AL levels.

3.3.2.2 Self-reported learning. Participants were asked to rate their learning level before and after CSA using a 5-point Likert scale (5 = the highest and 1 = the lowest) (Table 2).

3.3.2.3 Performance-based assessment tasks. The CSA Executive Committee designed and validated three performance-based tasks to objectively measure the trainees' workplace-related assessment skills. A scoring rubric was developed and validated for each task, and scorers were trained to improve the scoring and feedback quality (van de Ridder et al., 2020). This was used as a self-assessment tool to monitor the trainees' performance before submitting the task to the BB. Two attempts were made, followed by prompt feedback.

Task#1: Trainees were asked to apply knowledge and skills to develop and validate an analytic grading rubric for assessing learning domains.

Task#2: Trainees were asked to apply knowledge and skills to develop and validate type-A MCQs for assessing higher-order thinking skills.

Final Project: Trainees were asked to propose an intervention plan for SA practices at participants' universities (real-world context) based on an analysis of current and best practices at other universities, either by proposing new practices, improving existing practices (Cohort I), or developing new alternative assessment methods (Cohort II).

3.4 Procedure
3.4.1 Data collection. Data were collected by evaluating CSA assessments from 2020 to 2022. After each session, participants received a QR-code and link via PowerPoint presentation slides during CSA, via BB, or via email later. Figure 2 illustrates the participant assessment plan as illustrated in the CSA handbook (Centre for Quality of Assessment and Examinations, 2021).

3.4.2 Data analysis. Analyses were performed using IBM SPSS Statistics (version 26) (IBM Corp, 2019) and Amos (Version 26.0) (Arbuckle, 2019). Descriptive statistics (mean, standard deviation, and skewness) were used to examine data normality for each variable. This verified whether the statistical method (t-test and regression assumptions) was appropriate for hypothesis testing. If the t-test assumptions were violated, non-parametric tests, such as the Kolmogorov–Smirnov test and the Wilcoxon signed-rank test were used.

4. Results
To test H1, descriptive statistics for participants' responses before and after training sessions – mean, standard deviation, and skewness as computed by Pearson and Fischer coefficients (Hair et al., 2010) and percentages of respondents at each satisfaction level – were performed (Table 2). Participants’ responses were not normally distributed in post-training satisfaction. Approximately 85% of the participants assessed their satisfaction level as either 4 or 5 for each item and overall. The ratings ranged from 4.6 to 4.7 for all items. Moreover, the skewness values for post-training satisfaction reflect a non-normal data distribution and imply a high satisfaction level. As the non-normality of the post-training satisfaction data violated the assumptions of the t-test for one sample, a non-parametric Kolmogorov–Smirnov one-sample
A test was conducted to examine the sample data fit to a theoretical distribution. The static KS test ($D(710) = 0.437, p<.01$) showed that participants’ satisfaction significantly deviated from the theoretical distribution, as represented by the baseline/reference of “1,” the starting point of the satisfaction scale. Hence, the results verified H1.
H2a was tested in a similar way (Table 3). Approximately 85% of the participants rated their learning level as 4 or 5, indicating higher post-training learning levels. Furthermore, the mean of the learning levels before and after training was 3.8 and 4.6, respectively. Table 3 shows skewed data after training, which violates the assumptions of the $t$-test. Therefore, the Wilcoxon signed-rank test was conducted to compare the participants’ self-reported learning before and after attending CSA. On average, participants reported lower learning levels pre-training (mean = 4) than post-training (mean = 5), demonstrating a statistically significant difference ($T = 86,444.5$, $z = -17.1$, $p < 0.01$).

To test H2b, descriptive statistics for participants’ responses (mean, standard deviation, and skewness as computed by Pearson and Fischer coefficients) for pre- and post-training learning levels were calculated using subjective (Table 2) and objective measures (Table 3). As the non-normality of the post-training satisfaction data violated the paired-samples $t$-test assumptions, a non-parametric Wilcoxon signed-rank test was conducted to compare participants’ self-reported learning performance pre- and post-training (a subjective measure). On average, participants reported learning less before participating (Mdn = 3.83) than after (Mdn = 4.58), showing a statistically significant difference ($T = 3655.5$, $z = -17.13$, $p < 0.001$). Additionally, the differences in participants’ ALL scores (objective measure) were examined pre- and post-training. Table 3 shows a normal data distribution for the two variables, as indicated by the Pearson and Fischer values. Accordingly, a paired-sample $t$-test was conducted to compare participants’ learning pre- and post-training, as measured by the ALI. The results showed that participants demonstrated higher levels of learning after the training ($M = 64.28$; SD = 13.56) than before ($M = 52.79$; SD = 12.88) ($t[110] = -8.92$, $p < 0.01$). Descriptive statistics of the performance-based tasks (Table 2) showed that the high participants’ means exceeded 70%. Recorded scores for the post-training were 82.93, 86.55, and 81.87 for tasks 1, 2, and the final project, respectively. Hence, the results verified H2.

To test H3, the regression assumptions were first examined using multiple linear regression analysis (Abdelrahman, 2022). Standardized residuals were analyzed to check for outliers, but none were found (Std. Residual (Standardized Residual) min = −2.42, Std. Residuals; maximum = 1.82). Multicollinearity was assessed using tolerance and variance inflation factor (VIF). In addition, the data met the assumption of collinearity, indicating that it was not a concern (Gender, Tolerance = 0.97, VIF = 1.03; educational qualification, Tolerance = 0.96, VIF = 1.05; academic cluster, Tolerance = 0.98, VIF = 1.02). The data also satisfied the assumption of independent errors (Durbin–Watson value = 2.11). Finally, the histogram of standardized residuals indicated that the data contained approximately normally distributed errors, similar to the plot of the normally predicted probability (P-P) of the regression residuals of the dependent variable (assessment), which was consistent with the diagonal line in the plot. Therefore, multiple regressions were performed to examine

<table>
<thead>
<tr>
<th>Learning measure</th>
<th>M</th>
<th>SD</th>
<th>Skew (Pearson, Fischer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALI Pre-Assessment Score</td>
<td>52.8</td>
<td>13.6</td>
<td>−0.03, −0.43</td>
</tr>
<tr>
<td>ALI Post-Assessment Score</td>
<td>64.3</td>
<td>12.9</td>
<td>−0.18, −0.25</td>
</tr>
<tr>
<td>Task#1 Analytic scoring rubric development score</td>
<td>82.9</td>
<td>9.5</td>
<td>−0.09, −1.89</td>
</tr>
<tr>
<td>Task#2 Developing MCQs for assessing HOTs score</td>
<td>86.6</td>
<td>10.1</td>
<td>−0.14, −2.58</td>
</tr>
<tr>
<td>Final project score</td>
<td>82</td>
<td>8.8</td>
<td>0.003, −1.51</td>
</tr>
</tbody>
</table>

**Note(s):** ALI: assessment literacy inventory, MCQs: multiple choice questions, HOTs: higher order thinking skills

**Source(s):** The author’s own work

---

Table 3. Descriptive statistics of learning level measures ($n = 111$)
whether gender, academic cluster, and educational qualification predicted the total ALI (Table 4). The forced method revealed that gender, academic cluster, and educational qualification had significant variance in the total postsecondary learning levels of AL. The model explained 8.3% of the variance ($F[3, 107] = 3.32; p < 0.01; R^2 = 0.083$). Further analysis of the individual predictors showed that only the academic cluster was a significant predictor of the post-training learning levels of AL in the model ($b = -0.27; p < 0.01$).

5. Discussion
HEIs evaluate the effectiveness of FD programs to verify if they are worth the investment. This study evaluates the impact of faculty education via an FD program (CSA) using the Kirkpatrick model. Overall, CSA’s positive impact on educators can be attributed to its long-term format, which demonstrates the highest investment (Daumiller et al., 2021). The results indicated that CSA positively affects faculty members in the Saudi higher education context, as demonstrated the subjective and objective measurements of their satisfaction and learning levels. Overall, CSA’s positive impact was consistent with the benefits and outcomes of FD programs previously reported in teaching practices (e.g. Condon et al., 2016), SA practices (e.g. Hassanien, 2018), student learning (e.g. Daumiller et al., 2021); and receipt of programmatic accreditation (e.g. Alkathiri, 2020). These findings support the view that FD is an imperative institutional strategy for boosting teachers’ quality in response to rapid changes and demands in the labor market (Muammar and Alkathiri, 2021). Generally, the findings showed that the Kirkpatrick model maintains its adaptability in FD training contexts in various disciplines (Alsalamah and Callinan, 2021), such as health (Johnston et al., 2018) and social sciences (Mahmoodi et al., 2019), and can be extended to engineering, science, and management. Furthermore, the study findings showing elevated trainers’ satisfaction with CSA (Level 1/Response), consistent with results of previous studies assessing FD programs (Ghasemi et al., 2020; Johnston et al., 2018; Shah et al., 2023). While Steinert and Mann (2006) question the value of using satisfaction, which is merely an indicator of “happiness,” other researchers believe that assessing satisfaction using well-designed instruments yields valuable feedback for program planners. It also provides opportunities to improve FD programs by examining the factors contributing to overall faculty satisfaction and program success, such as realistic objectives, appropriate content, facilitators, and delivery (Muammar and Alkathiri, 2021). This is also applicable to our results, showing that considering faculty needs while developing CSA is expected to leverage participants’ motivation and result in positive outcomes. Moreover, the results showed learning gains as measured by subjective and objective measures, which confirm the CSA’s cost-effectiveness and the visibility of its investment in human capital, as previously reported for FD programs (e.g. Muammar and Alkathiri, 2021). This finding is consistent with Kirkpatrick’s (1996) argument that high satisfaction levels do not ensure that training objectives will be fulfilled, as that requires assessing levels beyond the trainees’ reaction. Senel (2014) also argued that using criterion-referenced performance tasks

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta(β)</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1.92</td>
<td>2.42</td>
<td>0.075</td>
<td>0.429</td>
</tr>
<tr>
<td>Academic cluster</td>
<td>-4.43</td>
<td>1.49</td>
<td>-0.277</td>
<td>0.004</td>
</tr>
<tr>
<td>Educational qualification</td>
<td>1.71</td>
<td>3.10</td>
<td>0.052</td>
<td>0.583</td>
</tr>
</tbody>
</table>

Note(s): $B$: unstandardized beta, Std. Error: standard error, $\beta$: standardized beta, Sig: significance of $p$
Source(s): The author’s own work
(objective measures) to properly evaluate learning level is crucial to provide a standard for
developing trainees’ skills against the expected behavior level, particularly in “complex
jobs.” This is evidenced by the trainees’ high level of learning, with a mean score ranging
from 80% to 86% in the performance-based tasks. This finding demonstrates that CSA
programs can be accepted successfully because they result in the extensive acquisition of the
target skills and behaviors through learning (Senel, 2014).

Additionally, the results emphasized SA’s role in HEIs, including the short-term and long-
term benefits of objectively assessing performance. Driven by the competency-based
assessment approach, educating faculty on SA in general, and AL in particular, is widely
recognized as an essential teaching competence (Willoughby-Knox, 2019), and assessing
learning levels using performance-based tasks reveals the acquisition of job performance-
related skills. The findings confirmed FD programs’ usefulness in developing educators’
knowledge and fostering their PD, as evidenced by their performance on tasks that required
planning and developing assessment tools and methods (Mahmoodi et al., 2019).

Furthermore, the trainees demonstrated significant improvements over time in their post-
ALI scores compared with their pre-ALI scores due to FD programs or experience, as
previously shown by Ghasemi et al. (2020), Johnston et al. (2018), and Mertler (2003).

Additionally, the results highlighted the significant influence of academic cluster (major)
on post-ALI scores, which was more pronounced in the health cluster than in others
(humanities, engineering, and science). Teachers in health-related professions were found to
have more assessment knowledge as measured by ALI scores, which could be attributed to
knowledge acquired in their medical education. This result is consistent with Senel (2014),
showing the effect of the trainees’ educational background on their post-training performance
in learning objective measures. Furthermore, no significant effects of gender or educational
qualification on ALI scores were observed, as previously reported in the Saudi context
(Jawhar and Subahi, 2020). Conversely, other contexts showed a significant effect of gender
on ALI scores (Iqbal et al., 2023) and on objective learning measures (Senel, 2014). Saudi HEIs
exhibited similar practices manifested in SA with strictly regulated policies, general
guidelines, and assessment standards passed on to the faculty regardless of their gender,
qualification, or teaching experience. This ensures that when SA complies with international
best practices, the impact of other factors may be ruled out (Jawhar and Subahi, 2020). In
addition, assessing teachers’ AL in various contexts showed an unsurprisingly inadequate
AL level, regardless of educational level, contextual differences, or AL assessment tools, as
previously shown by Jawhar and Subahi (2020), Mertler (2003), and Plake et al. (1993).

Therefore, training on AL was recommended (e.g. Iqbal et al., 2023; Jawhar and Subahi, 2020;
Levy-Vered and Naser-Abu Alhija, 2015; Mertler, 2003), which was the motivation for
conducting the current study. Consistent with these studies, this study’s results showed an
inadequate AL level among faculty from different disciplines, which was significantly
improved by CSA. These results call for developing similar training programs focused on SA
and AL competencies, including reliability and validity standards. Accordingly, HEI
policymakers and authorities could consider the present study’s findings as a compass for
planning effective long-term training programs for educators that foster SA quality at
individual and institutional levels. The study highlights teachers’ high levels of satisfaction
and readiness to learn, which are expected to positively impact subsequent levels, as
evidenced by behavioral changes in work practices and student performance (Shah et al.,
2023). Therefore, theory and practice should be combined when developing intensive SA and
AL training programs to promote student success and increase self-efficacy (Mahmoodi et al.,
2019). Finally, as HEIs seek recognition for their T&L practices and strive to ensure
excellence in graduates, promoting faculty quality by providing high-quality developmental
programs is an inevitable strategy.
6. Conclusions and implications
This study examined the impact of SA skills training on faculty satisfaction and learning levels, following the Kirkpatrick model. The findings support the positive influence of a long-term FD program, namely CSA, with both objective and subjective measures showing elevated satisfaction levels and significant learning gains. Further, the study emphasized the significant predictive role of academic cluster in faculty scores, as measured by ALI.

The study findings have important theoretical and practical implications for HEIs, highlighting that they should consider developing long-term FD programs on SA and AL. Such initiatives guarantee the caliber of SA practices and enhance the validity and reliability of assessment results. This guarantee would foster confidence in the educational service offered by HEIs. Further, investment in human capital via FD programs on SA could be favorable, positively impacting learning outcomes, empowering graduates, and supporting national and international accreditation efforts.

As this study was confined to the Saudi context, its limitations must be acknowledged. The results’ generalizability may be restricted by the sample size and selection. Moreover, the relatively low reliability of ALI prompts further exploration. Subsequent research endeavors could explore alternative measures for assessing or calibrating AL using theoretical frameworks beyond classical test theory. Future studies can expand similar FD programs to new contexts and examine the mediating role of other contextual variables, such as SA practice experience. Moreover, investigating the macro-level effects of training educators in SA areas, including behavior change and return on investment – as expressed in metrics such as teacher performance and student learning outcomes – using objective measures, would contribute to the academic discourse.

Note
1. The author initiated a communication with Prof. Mertler on January 9, 2020, to obtain permission for developing a validated Arabic version of the inventory.

References


Centre for Quality of Assessment and Examinations (2021), *An Introductory Handbook to Certificate of Student Assessment Program*, Imam Abdul Rahman Bin Faisal University, Damam, KSA.


IBM Corp (2019), IBM SPSS Statistics for Windows Version 26.0, IBM Corp, Armonk, NY.


Senel, S. (2014), “Evaluation of a leadership training program by applying Kirkpatrick’s four-level model: the case of Turkish prefect interns training and development program”, [Doctoral thesis], The University of Texas at Dallas.


About the author
Dr Wessam Abdel Moaty Mohamed, Dr rer.nat, SFHEA recently obtained a senior fellowship in Advance HE, UK. She is Assistant Professor of Educational Psychology and the Head of the Measurement & Testing Unit at The Centre for Quality of Measurement and Examinations (QA Centre), the Office of Vice–Presidency for Academic Affairs, Imam Abdulrahman University (IAU), KSA. She has participated in the Student Assessment House of Expertise, IAU, since 2022 as cofounder member and trainer in Arabic and English. She is one of the members of the QA Centre who has made an instrumental contribution to the development of the Certificate of Student Assessment Program (CSA) since 2018, a member of the CSA executive committee, one of the CSA trainers and assessors, and the head of the assessment committee of the CSA program since 2020. One of the significant responsibilities of the Head of the Measurement and Testing Unit is to evaluate the impact of QA-provided training on student assessment practices in academic programs at the individual and institutional levels. Wessam Mohamed can be contacted at: waabdelmoaty@iau.edu.sa