Development and validation of the EDUcational Course Assessment TOOLkit (EDUCATOOL) — a 12-item questionnaire for evaluation of training and learning programmes

Tena Matolić[1], Danijel Jurakić[1]*, Zrinka Greblo Jurakić[2], Tošo Maršić[3] and Željko Pedišić[4]*

[1]Faculty of Kinesiology, University of Zagreb, Zagreb, Croatia, [2]Faculty of Croatian Studies, University of Zagreb, Zagreb, Croatia, [3]Faculty of Kinesiology, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia, [4]Institute for Health and Sport, Victoria University, Melbourne, VIC, Australia

Introduction: The instruments for evaluation of educational courses are often highly complex and specifically designed for a given type of training. Therefore, the aims of this study were to develop a simple and generic EDUcational Course Assessment TOOLkit (EDUCATOOL) and determine its measurement properties.

Methods: The development of EDUCATOOL encompassed: (1) a literature review; (2) drafting the questionnaire through open discussions between three researchers; (3) Delphi survey with five content experts; and (4) consultations with 20 end-users. A subsequent validity and reliability study involved 152 university students who participated in a short educational course. Immediately after the course and a week later, the participants completed the EDUCATOOL post-course questionnaire. Six weeks after the course and a week later, they completed the EDUCATOOL follow-up questionnaire. To establish the convergent validity of EDUCATOOL, the participants also completed the "Questionnaire for Professional Training Evaluation."

Results: The EDUCATOOL questionnaires include 12 items grouped into the following evaluation components: (1) reaction; (2) learning; (3) behavioural intent (post-course)/behaviour (follow-up); and (4) expected outcomes (post-course)/results (follow-up). In confirmatory factor analyses, comparative fit index (CFI = 0.99 and 1.00), root mean square error of approximation (RMSEA = 0.05 and 0.03), and standardised root mean square residual (SRMR = 0.07 and 0.03) indicated adequate goodness of fit for the proposed factor structure of the EDUCATOOL questionnaires. The intraclass correlation coefficients (ICCs) for convergent validity of the post-course and follow-up questionnaires were 0.71 (95% confidence interval [CI]: 0.61, 0.78) and 0.86 (95% CI: 0.78, 0.91), respectively. The internal consistency reliability of the evaluation components expressed using Cronbach's alpha ranged from 0.83 (95% CI: 0.78, 0.87) to 0.88 (95% CI: 0.84, 0.92) for the post-course questionnaire and from 0.95 (95% CI: 0.93, 0.96) to 0.97 (95% CI: 0.95, 0.98) for the follow-up questionnaire. The test−retest reliability ICCs for the overall evaluation scores of the post-course and follow-up questionnaires were 0.87 (95% CI: 0.78, 0.92) and 0.91 (95% CI: 0.85, 0.94), respectively.

Conclusion: The EDUCATOOL questionnaires have adequate factorial validity, convergent validity, internal consistency, and test−retest reliability and they can be used to evaluate training and learning programmes.

Introduction

Learning is one of the key components of daily time use across the world (Charmes, 2015). According to time-use surveys conducted in 37 countries, between 15 and 69% of adults aged 25–64 years participate in learning programmes (OECD, 2023). Training, learning, and educational courses and programmes (hereafter referred to as "educational courses") have multifaceted benefits for individuals and organisations (Kraiger, 2008). Educational courses are commonly developed to improve subject-specific knowledge, increase work productivity, promote healthy lifestyle, or encourage pro-environmental behaviours (Kahn et al., 2002; Arthur et al., 2003; McColgan et al., 2013; Cavallo et al., 2014; Hughes et al., 2016; Beinicke and Bipp, 2018; Dusch et al., 2018; Hauser et al., 2020).

Educational courses need to be evaluated, to determine their quality and potential areas of improvement (Wilkes and Bligh, 1999; Arthur et al., 2003; Kraiger, 2008). The recommended ways of evaluating educational courses have evolved over time (Bell et al., 2017), and they now involve complex processes necessitating the use of scientifically grounded and standardised methods (Guskey, 2000). For this purpose, over the past 80 years, various frameworks for the evaluation of educational courses have been developed (Tamkin et al., 2002; Moseley and Dessinger, 2009; Shelton, 2011; Stufflebeam, 2014; Perez-Soltero et al., 2019).

The Kirkpatrick's evaluation framework (Kirkpatrick and Kirkpatrick, 2006) is widely used to guide the assessment of educational courses, both in research and practice (Moreau, 2017). Its most recent version, "The New World Kirkpatrick model" (Kirkpatrick and Kirkpatrick, 2016), incorporates evaluation of participants' reactions to education, learning quality, behavioural change, and the effects/results of education.

The available instruments that can be used to evaluate educational courses based on Kirkpatrick's model are often highly complex and specifically designed for a given type of training (Kraiger, 2008; Thielsch and Hadzihalilovic, 2020). Therefore, their application may require a substantial amount of time while being limited in scope (Grohmann and Kauffeld, 2013). In addition, literature reviews have shown that educational course evaluation commonly focuses only on the first two "levels" of Kirkpatrick's framework, that is, reaction and learning (McColgan et al., 2013; Hughes et al., 2016; Reio et al., 2017). This is also supported by the data in the "*Association for Talent Development's report*" from 2016 where talent development professionals reported that reaction was evaluated in 88%, learning in 83%, behaviour in 60%, and results in 35% of their organisations (Ho, 2016). Possible reason for this is a lack of generic instruments that would be applicable to a wide spectrum of educational courses.

Therefore, the aims of this study were to: (1) develop a simple and generic questionnaire for the evaluation of educational courses by assessing respondents' reactions to education, learning quality, behavioural change, and the effects/results of education; and (2) determine its validity and reliability.

Materials and methods

Development of EDUCATOOL

The EDUcational Course Assessment TOOLkit (EDUCATOOL) was developed in four stages, from March to November 2021.

Literature review

In the first stage of EDUCATOOL development, we conducted a comprehensive literature review to identify existing conceptual frameworks and questionnaires used to evaluate educational courses. This included searches in five bibliographic databases: SPORTDiscus (through EBSCOHost), APA PsycInfo (through EBSCOHost), Web of Science core collection (including Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index – Social Science & Humanities, Book Citation Index – Social Sciences & Humanities), Google Scholar, and Scopus. Full-texts of 150 publications were reviewed, and findings from 40 relevant books and papers were summarised and considered before drafting the questionnaire (Supplementary File S1).

Drafting the questionnaire

Based on discussions guided by the literature review, in the second stage, three researchers (TM, ŽP, DJ) created the first draft of EDUCATOOL. The toolkit consisted of two complementary questionnaires (post-course and follow-up questionnaires) (Pedisic et al., 2023a), user guide (Pedisic et al., 2023a), and a Microsoft Excel spreadsheet for data cleaning and processing (i.e., EDUCATOOL calculator) (Pedisic et al., 2023b). The post-course questionnaire was designed to capture participants' immediate feedback, and it is meant to be administered immediately upon the completion of the educational course. The follow-up questionnaire was designed to evaluate longer-term impacts of the course, and it is meant to be administered preferably 1–6 months after completing the course.

Delphi survey with content experts

The Delphi method — a systematic, iterative process aimed at achieving expert consensus — was used in the third stage of questionnaire development, to improve the initial version of EDUCATOOL. The Delphi panel included five experts in the following fields: (1) survey design and psychometrics; (2) evaluation of educational courses; (3) education and training; (4) psychology; and (5) English language. An independent researcher, who was not involved in the Delphi panel, served as a moderator of the process. Before each round of the survey, the moderator distributed anonymous questionnaire and supplementary files (i.e., EDUCATOOL instructions, questionnaires, and calculator) to the panel members. Between the survey rounds, the moderator carefully considered suggestions from the panel and modified the documents accordingly. Three rounds of Delphi survey were conducted, before achieving a

consensus among the experts on the purpose, content, and wording of EDUCATOOL.

Consultations with end-users

In the fourth stage, we initiated a consultative process aimed at further refinement of EDUCATOOL. The consultations involved 20 individuals, potential end-users of EDUCATOOL, including: (1) professionals involved in the development, delivery, and evaluation of educational courses; (2) educators in secondary and tertiary degree courses (3) researchers; and (4) managers of private businesses that conduct educational courses. The potential end-users were asked to review the EDUCATOOL questionnaires, instructions, and calculator and provide suggestions on how to improve them. Based on their feedback, we made final modifications to the documents.

Assessing reliability and validity of EDUCATOOL

Study design

To simulate a scenario in which individuals attend an educational course and then evaluate it using EDUCATOOL, we asked the participants in our study to engage in the Sports Club for Health (SCforH) online course (Jurakic et al., 2021). The topic of SCforH online course is how to improve the quality and availability of health-enhancing sports programmes through sports clubs and associations. The course consists of seven units, including videos, interactive infographics, and quizzes. It usually takes between 20 and 30 min to complete the course. The SCforH online course has been included in the curriculum of several tertiary degree courses in Europe.

In October 2022, the participants completed the SCforH online course. Immediately after the course, they completed the EDUCATOOL post-course questionnaire. One week later, the post-course questionnaire was re-administered to participants to enable evaluating its test–retest reliability. Six weeks after the course, the participants completed the EDUCATOOL follow-up questionnaire. A week later, the participants were asked to complete the follow-up questionnaire again, to enable assessing its test–retest reliability. On all four survey occasions, the participants were also asked to complete the "Questionnaire for Professional Training Evaluation" (Grohmann and Kauffeld, 2013), to enable evaluation of convergent validity of EDUCATOOL post-course and follow-up questionnaires.

Participants

We invited all third-year students from the Faculty of Kinesiology, University of Zagreb, Croatia to participate in the study. They were selected purposefully as the study population, because the SCforH online course is intended for the current and future stakeholders in the sports sector and it is one of the learning topics at the third year of Master's of Kinesiology programme at the University of Zagreb. Our goal was to include at least 90 participants in the sample, to ensure a satisfactory width of the 95% confidence interval (CI) of the intraclass correlation coefficient (ICC ± 0.075), assuming an ICC of 0.80, according to the Bonnett's calculation (Bonett, 2002). The final sample consisted of 152 participants. Prior to participation in the study, all participants provided an informed consent. Through the consent form, the participants were informed that: (1) the participation in the survey is voluntary; (2) they are not required to respond to all

questions; (3) they may withdraw from the study at any time without providing a reason for withdrawal and without any consequences; (4) we will not collect any personal information other than their email address; (5) their individual responses will be kept confidential; and (6) the collected data will only be used for research purposes and published collectively, that is, as a summary of responses from all participants. The study protocol was approved by the Ethics Committee of the Faculty of Kinesiology, University of Zagreb (number: 10/2021).

Measures

The EDUCATOOL post-course and follow-up questionnaires included 12 items each, asking about: (1) satisfaction with the course; (2) relevance / usefulness of the course; (3) level of engagement in the course; (4) acquisition of new knowledge through the course; (5) retention of knowledge acquired through the course; (6) development of new skills through the course; (7) retention of skills that were developed through the course; (8) increase in the interest in the subject of the course; (9) use of the knowledge acquired in the course; (10) use of the skills developed in the course; (11) improvements in personal performance; and (12) wider benefits of the course. The items were grouped into the following evaluation components: (1) reaction (items 1–3); (2) learning (items 4–8); (3) behavioural intent (post-course)/behaviour (follow-up; items 9–10); and (4) expected outcomes (post-course)/results (follow-up; items 11–12). All items (i.e., statements) in the questionnaire were positive, to avoid possible issue with double negation in responses.

The Questionnaire for Professional Training Evaluation included 12 items asking about six factors (i.e., satisfaction, utility, knowledge, application to practice, individual results, and global results) grouped into four evaluation components: reaction; learning; behaviour; and organisational results. Details about the questionnaire can be found elsewhere (Grohmann and Kauffeld, 2013). Previous research has shown that the Questionnaire for Professional Training Evaluation has good discriminant validity and internal consistency reliability (Cronbach's $\alpha = 0.79$ to 0.96) (Grohmann and Kauffeld, 2013) For the purpose of this study, we slightly modified the original wording of the items, so that the questionnaire can be administered immediately after the course.

In both questionnaires, participants were asked to provide their responses on an 11-point Likert scale ranging from 0 ("completely disagree") to 10 ("completely agree"). The evaluation component scores for both questionnaires were calculated as the arithmetic means of the respective questionnaire items, while the overall evaluation score was calculated as the arithmetic mean of evaluation components. The questionnaires were administered in English, because we were interested in the measurement properties of the original, English version of EDUCATOOL.

Data analysis

To evaluate the factorial validity of the proposed 4-factor model, we conducted a confirmatory factor analysis using weighted least squares means and variance adjusted estimation. This method has been proposed for ordinal Likert-type data and it does not assume normal distribution of data (Beauducel and Herzberg, 2006; Brown, 2015). The model fit was assessed based on the following fit indices: (i) the scaled chi-square test; (ii) the comparative fit index (CFI); (iii) the root mean square error of

approximation (RMSEA), and (iv) the standardised root mean square residual (SRMR). The chi-square test p-value of <0.05 was considered to indicate a lack of good fit (Bollen and Stine, 1992; Kline, 2023), while CFI ≥ 0.95 (Hu and Bentler, 1999), RMSEA ≤ 0.06 (Steiger, 2007), and SRMR ≤ 0.08 (Hu and Bentler, 1999) were considered to indicate adequate model fit. We also calculated factor loadings for all questionnaire items and assessed them against the conservative threshold of 0.60 (Matsunaga, 2010). The internal consistency reliability of evaluation components and overall score was expressed using the Cronbach's alpha coefficient and its 95% CI. Convergent validity and test–retest reliability were expressed using the two-way mixed model intraclass correlation coefficient, type [A, 1], case 3A according to McGraw and Wong (McGraw and Wong, 1996) (single measure, absolute agreement) and its 95% CI. The data were analysed using RStudio (version 2022.07.1, Build 554) (RStudio v2022.07, 2022) using the packages "lavaan" (Rosseel et al., 2023), "lavaanPlot" (Lishinski, 2022), "MVN" (Korkmaz et al., 2022), "energy" (Rizzo and Szekely, 2022), "psych" (Revelle, 2022), and "boot" (Canty and Ripley, 2021).

Results

The final version of EDUCATOOL

During the three rounds of Delphi process, 39 changes have been made to EDUCATOOL. At the end of the process, the Delphi panel has reached a complete consensus on its content. EDUCATOOL underwent additional 10 changes as part of the consultations with end-users, and its final version includes: post-course questionnaire (Pedisic et al., 2023a); follow-up questionnaire (Pedisic et al., 2023a); user manual (Pedisic et al., 2023a); and Microsoft Excel spreadsheet for data processing (Pedisic et al., 2023b).

Reaction

For the purpose of the current study, we defined reaction as the degree to which participants find the educational course satisfactory, relevant/useful, and engaging. In the EDUCATOOL questionnaires, satisfaction is assessed with the item *"Overall, I am satisfied with this course,"* relevance with *"I find this course useful"* (post-course questionnaire) or *"This course has been useful to me"* (follow-up questionnaire), and engagement with *"I was fully engaged in this course."*

Learning

For the purpose of the current study, we defined learning as the degree to which participants gain and retain knowledge, develop, and retain skills, and increase their interest in the subject as a result of attending the course. In the EDUCATOOL questionnaires, knowledge acquisition is assessed with the item *"I acquired new knowledge in this course,"* knowledge retention with *"I will be able to retain this knowledge over the long term"* (post-course questionnaire) or *"I still possess the knowledge I acquired in this course"* (follow-up questionnaire), skill development with *"This course helped me develop skills,"* skill retention with *"I will be able to retain these skills over the long term"* (post-course questionnaire) or *"I still possess the skills developed in this course "* (follow-up questionnaire), and attitude change with *"Taking this course increased my interest in the subject."*

Behavioural intent/behaviour

For the purpose of the current study, we defined behavioural intent and behaviour as the degree to which participants utilise or intend to utilise the knowledge/skills gained in the course. In the post-course questionnaire, utilisation is assessed with the items: *"I will use the knowledge acquired in this course"* and *"I will use the skills developed in this course."* In the follow-up questionnaire, the items are worded: *"I have used the knowledge acquired in this course"* and *"I have used the skills developed in this course."*

Expected outcomes/results

For the purpose of the current study, we defined expected outcomes and results as the degree to which participation in the course resulted in or is expected to result in improvement of personal performance and other benefits. In the post-course questionnaire, they are assessed with the items: *"Participation in this course will improve my performance (e.g., work performance, academic performance, task-specific performance)"* and *"My participation in this course will result in other benefits (e.g., benefits for my business, institution, or community),"* respectively. In the follow-up questionnaire, the wording of these items is: *"Participation in this course has improved my performance (e.g., work performance, academic performance, task-specific performance)"* and *"My participation in this course resulted in other benefits (e.g., benefits for my business, institution, or community)."*

Measurement properties of EDUCATOOL

Factorial and convergent validity

In the confirmatory factor analysis of the proposed model with four factors including: (1) reaction; (2) learning; (3) behavioural intent (post-course)/behaviour (follow-up); and (4) expected outcomes (post-course)/results (follow-up), all goodness of fit statistics except the scaled chi-square test indicated adequate fit for the EDUCATOOL post-course and follow-up questionnaires (Table 1). The factor loadings in the confirmatory factor analysis for all items were above the 0.60 threshold, ranging from 0.66 to 0.92 for the post-course questionnaire (Table 2) and from 0.87 to 0.98 (Table 3) for the follow-up questionnaire. Furthermore, when assessed against the Questionnaire for Professional Training Evaluation, the convergent validity of the post-course and follow-up questionnaire was 0.71 (95% CI: 0.61, 0.78) and 0.86 (95% CI: 0.78, 0.91), respectively.

Internal consistency and test–retest reliability

The internal consistency reliability of the EDUCATOOL evaluation components ranged from 0.83 to 0.88 for the post-course

TABLE 1 Goodness of fit statistics for a four-factor structure of the EDUCATOOL questionnaire items.

Goodness of fit	Post-course questionnaire	Follow-up questionnaire
χ^2 (p)*	71.53 (0.015)	97.52 (<0.001)
RMSEA[†]	0.05	0.03
SRMR[‡]	0.07	0.03
CFI[§]	0.99	1.00

*Scaled chi-square (p-value). [†]Robust root mean square error of approximation.
[‡]Standardised root mean square residual. [§]Robust comparative fit index.

TABLE 2 Factor loadings, internal consistency, and test–retest reliability of the EDUCATOOL post-course questionnaire.

	Factor loading*	Cronbach's α (95% CI)[†]	ICC (95% CI)[‡]
Questionnaire item			
(1) Overall, I am satisfied with this course.	0.82	–	0.72 (0.60, 0.81)
(2) I find this course useful.	0.82	–	0.64 (0.46, 0.76)
(3) I was fully engaged in this course.	0.74	–	0.68 (0.55, 0.77)
(4) I acquired new knowledge in this course.	0.67	–	0.70 (0.58, 0.79)
(5) I will be able to retain this knowledge over the long term.	0.66	–	0.55 (0.39, 0.67)
(6) This course helped me develop skills.	0.77	–	0.75 (0.65, 0.83)
(7) I will be able to retain these skills over the long term.	0.72	–	0.57 (0.42, 0.69)
(8) Taking this course increased my interest in the subject.	0.69	–	0.58 (0.41, 0.70)
(9) I will use the knowledge acquired in this course.	0.87	–	0.77 (0.67, 0.84)
(10) I will use the skills developed in this course.	0.88	–	0.69 (0.56, 0.78)
(11) Participation in this course will improve my performance.	0.92	–	0.69 (0.57, 0.78)
(12) My participation in this course will result in other benefits.	0.87	–	0.66 (0.53, 0.76)
Evaluation component			
Reaction	–	0.84 (0.78, 0.88)	0.74 (0.61, 0.83)
Learning	–	0.83 (0.78, 0.87)	0.81 (0.72, 0.87)
Behavioural intent	–	0.87 (0.81, 0.91)	0.78 (0.68, 0.85)
Expected outcomes	–	0.88 (0.84, 0.92)	0.73 (0.62, 0.81)
Overall evaluation score	–	0.93 (0.91, 0.94)	0.87 (0.78, 0.92)

*Factor loadings on Reaction (items 1–3), Learning (items 4–8), Behavioural intent (items 9–10), and Expected outcomes (items 11–12) from the confirmatory factor analysis. [†]Internal consistency reliability expressed using Cronbach's alpha coefficient and its 95% confidence interval. [‡]One-week test–retest reliability expressed using intraclass correlation coefficient type (A,1) case 3A, according to McGraw and Wong (1996) and its 95% confidence interval.

questionnaire and from 0.95 to 0.97 for the follow-up questionnaire. The internal consistency reliability of the overall evaluation score from the post-course and follow-up questionnaires was 0.93 and 0.98, respectively (Tables 2, 3).

The test–retest reliability of the EDUCATOOL post-course questionnaire items ranged from 0.55 (95% CI: 0.39, 0.67) for *knowledge retention* ("I will be able to retain this knowledge over the long term") to 0.77 (95% CI: 0.67, 0.84) for *knowledge utilisation* ("I will use the knowledge acquired in this course"; Table 2). The test–retest reliability of evaluation components ranged from 0.73 (95% CI: 0.62, 0.81) for *expected outcomes* to 0.81 (95% CI: 0.72, 0.87) for *learning*. The test–retest reliability of the overall evaluation score was 0.87 (95% CI: 0.78, 0.92).

The test–retest reliability of the EDUCATOOL follow-up questionnaire items ranged from 0.75 (95% CI: 0.63, 0.83) for *satisfaction* ("Overall, I am satisfied with this course") and *skill retention* ("I still possess the skills developed in this course") to 0.85 (95% CI: 0.77, 0.90) for a*ttitude change* ("Taking this course increased my interest in the subject"; Table 3). The test–retest reliability of evaluation components ranged from 0.80 (95% CI: 0.70, 0.87) for *reaction* to 0.88 (95% CI: 0.82, 0.93) for *learning*. The test–retest reliability of the overall evaluation score was 0.91 (95% CI: 0.85, 0.94).

Discussion

Key findings

The literature review, open discussions between three researchers, Delphi survey with five content experts, and consultations with 20

end-users have informed the development of the EDUCATOOL post-course and follow-up questionnaires. These 12-item questionnaires can be used to evaluate training and learning programmes through the assessment of participants' reaction, learning, behavioural intent/ behaviour, and expected outcomes/results.

The key finding of this study is that the EDUCATOOL questionnaires have good measurement properties. In specific, our confirmatory factor analyses found a good fit for the proposed factor structure of EDUCATOOL questionnaire items. For both EDUCATOOL questionnaires, we also found adequate convergent validity, internal consistency, and test–retest reliability.

Factorial and convergent validity

Our analyses have confirmed the hypothesised 4-factor structure of EDUCATOOL questionnaire items. The number of factors is in accordance with the Kirkpatrick's evaluation framework (Kirkpatrick and Kirkpatrick, 2006, 2016) that is widely used as a guide for the assessment of educational courses, and with the factor structure of some previous questionnaires in this field (Cassel, 1971; Johnston et al., 2003). In comparison, a previous study found a six-factor structure of the Questionnaire for Professional Training Evaluation, with the factors representing participant satisfaction, perceived utility, gained knowledge, application to practice, individual organisational results, and global organisational results (Grohmann and Kauffeld, 2013). The difference between the two questionnaires in the factor structure is likely due to the differences in the wording and content of their items. For example, unlike the Questionnaire for Professional

TABLE 3 Factor loadings, internal consistency, and test–retest reliability of the EDUCATOOL follow-up questionnaire.

	Factor loading*	Cronbach's α (95% CI)†	ICC (95% CI)‡
Questionnaire item			
(1) Overall, I am satisfied with this course.	0.92	–	0.75 (0.63, 0.83)
(2) This course has been useful to me.	0.98	–	0.81 (0.71, 0.88)
(3) I was fully engaged in this course.	0.88	–	0.76 (0.64, 0.84)
(4) I acquired new knowledge in this course.	0.87	–	0.77 (0.66, 0.85)
(5) I still possess the knowledge I acquired in this course.	0.92	–	0.84 (0.75, 0.89)
(6) This course helped me develop skills.	0.93	–	0.81 (0.71, 0.87)
(7) I still possess the skills developed in this course.	0.92	–	0.75 (0.63, 0.83)
(8) Taking this course increased my interest in the subject.	0.92	–	0.85 (0.77, 0.90)
(9) I have used the knowledge acquired in this course.	0.95	–	0.81 (0.70, 0.88)
(10) I have used the skills developed in this course.	0.98	–	0.76 (0.64, 0.85)
(11) Participation in this course has improved my performance.	0.98	–	0.78 (0.66, 0.86)
(12) My participation in this course resulted in other benefits.	0.95	–	0.79 (0.68, 0.86)
Evaluation component			
Reaction	–	0.95 (0.93, 0.96)	0.80 (0.70, 0.87)
Learning	–	0.96 (0.95, 0.97)	0.88 (0.82, 0.93)
Behaviour	–	0.97 (0.95, 0.98)	0.81 (0.69, 0.88)
Results	–	0.96 (0.95, 0.97)	0.81 (0.70, 0.88)
Overall evaluation score	–	0.98 (0.97, 0.98)	0.91 (0.85, 0.94)

*Factor loadings on Reaction (items 1–3), Learning (items 4–8), Behaviour (items 9–10), and Results (items 11–12) from the confirmatory factor analysis. †Internal consistency reliability expressed using Cronbach's alpha coefficient and its 95% confidence interval. ‡One-week test–retest reliability expressed using intraclass correlation coefficient type (A,1) case 3A, according to McGraw and Wong (1996) and its 95% confidence interval.

Training Evaluation, the EDUCATOOL questionnaires ask about the engagement in the course, skill development and utilisation, knowledge and skill retention, and attitude change.

Despite these differences, the convergent validity of EDUCATOOL established against the Questionnaire for Professional Training Evaluation is relatively high, indicating that the questionnaires assess a similar construct. The convergent validity was higher for the follow-up questionnaire, compared with the post-course questionnaire, which may be attributed to the fact that the original version of the Questionnaire for Professional Training Evaluation is intended to be administered at least 4 weeks after the educational course. In comparison, the convergent validity of the FIRE-B questionnaire (Thielsch and Hadzihalilovic, 2020), that was developed based on the Kirkpatrick's evaluation framework, was somewhat lower than for EDUCATOOL, ranging from 0.45 to 0.69.

Internal consistency and test–retest reliability

Both EDUCATOOL questionnaires have adequate internal consistency and test–retest reliability, comparable with other questionnaires for course evaluation (Aleamoni and Spencer, 1973; Byrne and Flood, 2003; Royal et al., 2018; Niemann and Thielsch, 2020). The test–retest reliability varied across EDUCATOOL questionnaire items, with the lowest (albeit still satisfactory) ICCs found for the items on knowledge retention,

skills retention, and attitude change in the post-course questionnaire. It is possible that some participants overestimated or underestimated their knowledge/skills retention and attitude change immediately after the course (i.e., at the time of the first survey), while they were able to estimate it more accurately a week later (i.e., at the time of the re-test survey). This possible explanation is supported by the fact that the respective questions in the follow-up survey have somewhat higher test–retest reliability. This explanation is also supported by previous findings on a relatively high level of participant knowledge immediately after the training, which then reduces over time (Ritzmann et al., 2014). Importantly, the resulting evaluation component (*learning*) from the EDUCATOOL post-course questionnaire seems to have a higher test–retest reliability (ICC = 0.81) than the belonging individual items.

In our study sample, the overall evaluation score, the four evaluation components, and all individual items of the EDUCATOOL follow-up questionnaire have shown somewhat higher test–retest reliability, compared with the post-course questionnaire. It is possible that the outcomes of course attendance stabilise over time, making participants more likely to respond to the questionnaire in a consistent manner. It could also be that the follow-up questionnaire captures more stable aspects of educational experience which are less likely to change over time. These possible explanations are in accordance with the findings of previous methodological studies indicating that the questions about the past generally have higher reliability than the

questions pertaining to the present and future (Tourangeau, 2021). The overall evaluation score and four evaluation components of the EDUCATOOL follow-up questionnaire also seem to have somewhat higher internal consistency reliability, compared with the post-course questionnaire.

Implications for research and practice

The generic wording of EDUCATOOL questionnaire items will enable its use for the evaluation of different types of educational courses (e.g., online or face-to-face, professional or recreational, long or short) across various fields and settings. An additional advantage of EDUCATOOL is its brevity, making it a practical choice for collecting valuable course evaluation data even in situations with limited time available. While EDUCATOOL can provide a good insight into participants' reactions to education, learning quality, behavioural change, and the effects/results of education, for a more comprehensive evaluation, the use of additional methods and evaluation tools may need to be considered. For example, researchers and practitioners may find it relevant to examine different types of interactions in the learning process (Moore, 1989), instructor's effectiveness (Kuo et al., 2014), transfer of learning (Blume et al., 2010), and monetary benefits of course attendance (Phillips and Phillips, 2016), which cannot be assessed directly or in detail using EDUCATOOL.

Strengths and limitations of the study

Our study had the following strengths: (1) a systematic approach used to inform the development of EDUCATOOL; (2) a diverse group of experts involved in the Delphi panel; (3) a large number of potential end-users of the questionnaire who have contributed to the consultation process; and (4) a relatively large number of participants involved in the study of validity and reliability.

Our study had several limitations. First, the study was conducted in a convenience sample, limiting the generalisability of our findings. Future studies should examine measurement properties of EDUCATOOL in representative samples of various population groups, such as students from various colleges. Second, due to the differences in the factor structure of EDUCATOOL and the Questionnaire for Professional Training Evaluation, in this study we were only able to examine the convergent validity of the overall evaluation score. Future studies should consider exploring the convergent validity of EDUCATOOL also against other questionnaires for evaluation of educational courses. Third, in the study of validity and reliability, the EDUCATOOL questionnaire referred to a single online course; thus, it would be beneficial to further investigate the application of EDUCATOOL in other training areas and with other types of courses. Fourth, the EDUCATOOL questionnaire used in this study was in English and the participants were non-native English speakers. Despite the fact that all participants in our sample had at least 9 years of formal education in English as secondary language, it might be that the measurement

properties of EDUCATOOL would be somewhat different if the study was conducted among native English speakers.

Conclusion

The EDUCATOOL post-course and follow-up questionnaires can be used to evaluate training and learning programmes through the assessment of participants' reaction, learning, behavioural intent/behaviour, and expected outcomes/results. The novel questionnaires have adequate factorial validity, convergent validity, internal consistency, and test–retest reliability. Given the generic wording of their items, the questionnaires can be used to evaluate different types of courses in various fields. Future studies should examine measurement properties of EDUCATOOL in representative samples of different population groups attending various courses.

Data availability statement

The datasets presented in this article are not readily available because they are anonymous and intended for study purposes only. Requests to access the datasets should be directed to TM, tena.matolic@kif.unizg.hr.

Ethics statement

The study was approved by the Ethics Committee of the Faculty of Kinesiology, University of Zagreb (number: 10/2021). The study was conducted in accordance with the local legislation and institutional requirements. The participants provided their informed consent to participate in this study.

Author contributions

TeM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Writing – original draft, Writing – review & editing. DJ: Conceptualization, Methodology, Software, Supervision, Validation, Writing – review & editing. ZGJ: Software, Validation, Writing – review & editing. ToM: Software, Validation, Writing – review & editing. ŽP: Conceptualization, Methodology, Software, Supervision, Validation, Writing – review & editing.

Funding

Researchers' Career Development Project, funded by Croatian Science Foundation (ref: DOK-2020-01-8078).

Acknowledgments

The authors wish to express their gratefulness to all participants in the Delphi panel, consultations process, and validation study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union, The European Education and Culture Executive Agency (EACEA), or the Croatian Science Foundation. Neither the European Union nor the granting authority can be held responsible for them.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1314584/full#supplementary-material

References

Aleamoni, L. M., and Spencer, R. E. (1973). The Illinois course evaluation questionnaire: a description of its development and a report of some of its results. *Educ. Psychol. Meas.* 33, 669–684. doi: 10.1177/001316447303300316

Arthur, W. Jr., Bennett, W. Jr., Edens, P. S., and Bell, S. T. (2003). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *J. Appl. Psychol.* 88, 234–245. doi: 10.1037/0021-9010.88.2.234

Beauducel, A., and Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Struct. Equ. Model.* 13, 186–203. doi: 10.1207/s15328007sem1302_2

Beinicke, A., and Bipp, T. (2018). Evaluating training outcomes in corporate e-learning and classroom training. *Vocat. Learn.* 11, 501–528. doi: 10.1007/s12186-018-9201-7

Bell, B., Tannenbaum, S., Ford, J., Noe, R., and Kraiger, K. (2017). 100 years of training and development research: what we know and where we should go. *J. Appl. Psychol.* 102, 305–323. doi: 10.1037/apl0000142

Blume, B. D., Ford, J. K., Baldwin, T. T., and Huang, J. L. (2010). Transfer of training: a meta-analytic review. *J. Manage.* 36, 1065–1105. doi: 10.1177/0149206309352880

Bollen, K. A., and Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociol. Methods Res.* 21, 205–229. doi: 10.1177/004912419202100200

Bonett, D. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Stat. Med.* 21, 1331–1335. doi: 10.1002/sim.1108

Brown, TA. *2nd Confirmatory factor analysis for applied research*. New York, NY: Guilford Press (2015).

Byrne, M., and Flood, B. (2003). Assessing the teaching quality of accounting programmes: an evaluation of the course experience questionnaire. *Assess. Eval. High. Educ.* 28, 135–145. doi: 10.1080/02602930301668

Canty, A, and Ripley, B. *Package 'boot'*. Vienna: The Comprehensive R Archive Network (2021).

Cassel, R. N. (1971). A Student Course Evaluation Questionnaire. *Improving Coll. Univ. Teach.* 19, 204–206. doi: 10.1080/00193089.1971.10533113

Cavallo, D. N., Brown, J. D., Tate, D. F., DeVellis, R. F., Zimmer, C., and Ammerman, A. S. (2014). The role of companionship, esteem, and informational support in explaining physical activity among young women in an online social network intervention. *J. Behav. Med.* 37, 955–966. doi: 10.1007/s10865-013-9534-5

Charmes, J. *Time use across the world: Findings of a world compilation of time use surveys*. New York, NY: United Nations Development Programme. (2015).

Dusch, M., Narciß, E., Strohmer, R., and Schüttpelz-Brauns, K. (2018). Competency-based learning in an ambulatory care setting: implementation of simulation training in the ambulatory care rotation during the final year of the MaReCuM model curriculum. *GMS J. Med. Educ.* 35, 1–23. doi: 10.3205/zma001153

Grohmann, A., and Kauffeld, S. (2013). Evaluating training programs: development and correlates of the questionnaire for professional training evaluation. *Int. J. Train. Dev.* 17, 135–155. doi: 10.1111/ijtd.12005

Guskey, TR. *Evaluating professional development*. Thousand Oaks, CA: Corwin Press. (2000).

Hauser, A., Weisweiler, S., and Frey, D. (2020). Because 'happy sheets' are not enough – a meta-analytical evaluation of a personnel development program in academia. *Stud. High. Educ.* 45, 55–70. doi: 10.1080/03075079.2018.1509306

Ho, M. *Evaluating learning: Getting to measurements that matter*. Alexandria, VA: ATD Press (2016).

Hu, L.-T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Modeling* 6, 1–55. doi: 10.1080/10705519909540118

Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., et al. (2016). Saving lives: a meta-analysis of team training in healthcare. *J. Appl. Psychol.* 101, 1266–1304. doi: 10.1037/apl0000120

Johnston, J. M., Leung, G. M., Fielding, R., Tin, K. Y., and Ho, L. M. (2003). The development and validation of a knowledge, attitude and behaviour questionnaire to assess undergraduate evidence-based practice teaching and learning. *Med. Educ.* 37, 992–1000. doi: 10.1046/j.1365-2923.2003.01678.x

Jurakic, D, Matolic, T, Bělka, J, Benedicic Tomat, S, Broms, L, De Grauwe, G, et al. *Sports Club for Health (SCforH) online course*. Zagreb: SCforH Consortium. (2021)

Kahn, E. B., Ramsey, L. T., Brownson, R. C., Heath, G. W., Howze, E. H., Powell, K. E., et al. (2002). The effectiveness of interventions to increase physical activity: a systematic review. *Am. J. Prev. Med.* 22, 73–107. doi: 10.1016/s0749-3797(02)00434-8

Kirkpatrick, D, and Kirkpatrick, J. *Evaluating training programs: The four levels*. Oakland, CA: Berrett-Koehler Publishers (2006)

Kirkpatrick, JD, and Kirkpatrick, WK. *Kirkpatrick's four levels of training evaluation*. Alexandria, VA: ATD Press (2016).

Kline, RB. *Principles and practice of structural equation modeling*. New York, NY: Guilford Press (2023).

Korkmaz, S, Goksuluk, D, and Zararsiz, G. *Package 'MVN'*. Vienna: The Comprehensive R Archive Network. (2022).

Kraiger, K. (2008). Benefits of training and development for individuals and teams, organizations, and society. *Annu. Rev. Psychol.* 60, 451–474. doi: 10.1146/annurev.psych.60.110707.163505

Kuo, Y.-C., Walker, A. E., Schroder, K. E. E., and Belland, B. R. (2014). Interaction, internet self-efficacy, and self-regulated learning as predictors of student satisfaction in online education courses. *Internet High. Educ.* 20, 35–50. doi: 10.1016/J.IHEDUC.2013.10.001

Lishinski, A. *Package 'lavaanPlot'*. Vienna: The Comprehensive R Archive Network. (2022).

Matsunaga, M. (2010). How to factor-analyze your data right: do's, don'ts, and how-to's. *Int. J. Psychol. Res.* 3, 97–110. doi: 10.21500/20112084.854

McColgan, P., McKeown, P. P., Selai, C., Doherty-Allan, R., and McCarron, M. O. (2013). Educational interventions in neurology: a comprehensive systematic review. *Eur. Neurol.* 20, 1006–1016. doi: 10.1111/ene.12144

McGraw, K. O., and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* 1, 30–46. doi: 10.1037/1082-989X.1.1.30

Moore, M. (1989). Three types of interaction. *Am. J. Distance Educ.* 3, 1–7. doi: 10.1080/08923648909526659

Moreau, K. A. (2017). Has the new Kirkpatrick generation built a better hammer for our evaluation toolbox? *Med. Teach.* 39, 1–3. doi: 10.1080/0142159X.2017.1337874

Moseley, JL, and Dessinger, JC. *Handbook of improving performance in the workplace: measurement and evaluation*. San Francisco, CA: Pfeiffer (2009). *3*

Niemann, L., and Thielsch, M. T. (2020). Evaluation of basic trainings for rescue forces. *J. Homel. Secur. Emerg.* 17, 1–33. doi: 10.1515/jhsem-2019-0062

OECD Stat. (2023). https://stats.oecd.org/index.aspx?queryid=120166 (Accessed August 26, 2023).

Pedisic, Z, Matolic, T, and Jurakic, D. EDUCATOOL instructions and questionnaires. (2023a) http://educatool.org/ (Accessed October 10, 2023).

Pedisic, Z, Matolic, T, and Jurakic, D. EDUCATOOL calculator. (2023b). Available at: http://educatool.org/ (Accessed October 10, 2023).

Perez-Soltero, A., Aguilar-Bernal, C., Barcelo-Valenzuela, M., Sanchez-Schmitz, G., Merono-Cerdan, A., and Fornes-Rivera, R. (2019). Knowledge transfer in training processes: towards an integrative evaluation model. *IUP J. Knowl. Manag.* 17, 7–40.

Phillips, JJ, Phillips, PP. *4th Handbook of training evaluation and measurement methods*. London: Routledge (2016)

Reio, T. G., Rocco, T. S., Smith, D. H., and Chang, E. (2017). A critique of Kirkpatrick's evaluation model. *New Horiz. Adult Educ.* 29, 35–53. doi: 10.1002/nha3.20178

Revelle, W. *Package 'psych'*. Vienna: The Comprehensive R Archive Network. (2022).

Ritzmann, S., Hagemann, V., and Kluge, A. (2014). The training evaluation inventory (TEI) - evaluation of training design and measurement of training outcomes for predicting training success. *Vocat. Learn.* 7, 41–73. doi: 10.1007/s12186-013-9106-4

Rizzo, M, and Szekely, G. *Package 'energy'*. Vienna: The Comprehensive R Archive Network. (2022).

Rosseel, Y, Jorgensen, D. T, Rockwood, N, Oberski, D, Byrnes, J, Vanbrabant, L, et al. *Package 'lavaan'*. Vienna: The Comprehensive R Archive Network. (2023)

Royal, K., Temple, L., Neel, J., and Nelson, L. (2018). Psychometric validation of a medical and health professions course evaluation questionnaire. *Am. Educ. Res. J.* 6, 38–42. doi: 10.12691/education-6-1-6

RStudio v2022.07. (2022) RStudio v2022.07. Available at: https://www.rstudio.com/products/rstudio/download/preview/

Shelton, K. (2011). A review of paradigms for evaluating the quality of online education programs. *Online J. Distance Learn. Edu.* 4, 1–11.

Steiger, J. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Pers. Individ. Differ.* 42, 893–898. doi: 10.1016/j.paid.2006.09.017

Stufflebeam, DL, , and Coryn, , CL. *2nd ed. Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass (2014).

Tamkin, P, Yarnall, J, and Kerrin, M. *Kirkpatrick and beyond: A review of models of training evaluation*. Brighton: Institute for Employment Studies (2002).

Thielsch, M. T., and Hadzihalilovic, D. (2020). Evaluation of fire service command unit trainings. *Int. J. Disaster Risk Sci.* 11, 300–315. doi: 10.1007/s13753-020-00279-6

Tourangeau, R. (2021). Survey reliability: models, methods, and findings. *J. Surv. Stat. Methodol.* 9, 961–991. doi: 10.1093/jssam/smaa021

Wilkes, M., and Bligh, J. (1999). Evaluating educational interventions. *Brit. Med. J.* 318, 1269–1272. doi: 10.1136/bmj.318.7193.1269