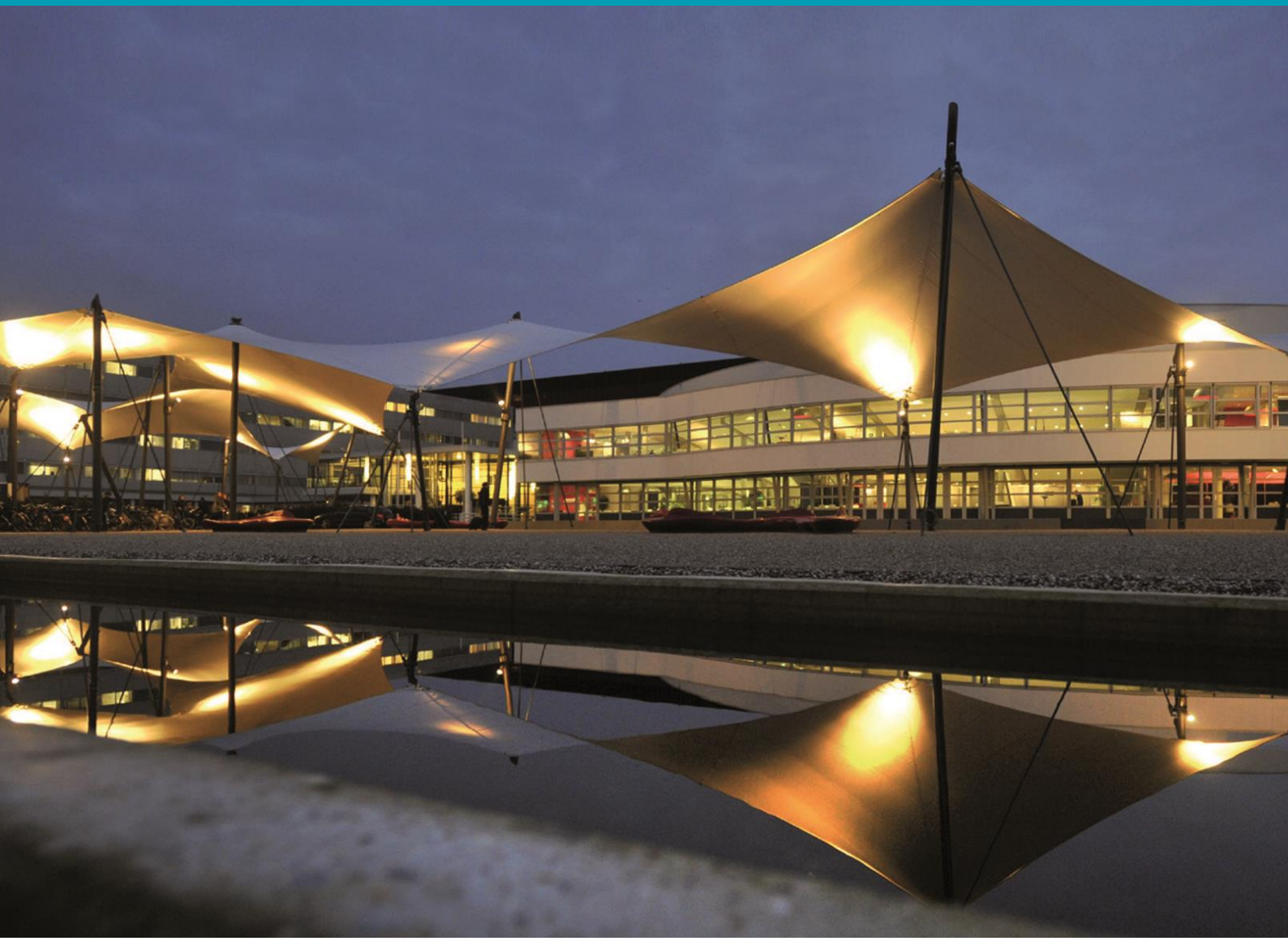
UNIVERSITY OF TWENTE.



Guiding the Evaluation of Upskilling Interventions in the Construction Industry

Leandro Hirose

Educational Science and Technology Master's Program

Faculty of Behavioural Sciences

UNIVERSITY OF TWENTE

June 2023

1st Supervisor

Dr. Sebastian M. Dennerlein

2nd Supervisor

Prof. Dr. Maaike Endedijk

Keywords: evaluation, assessment, upskill, construction industry, Kirkpatrick model

Word count: 21577

Acknowledgement

Guiding the design of evaluations in upskilling at a European level project was a complex task. Gladly I received the guidance of multiple people to successfully achieve this challenge. First, I would like to thank my supervisors, Dr Sebastian M. Dennerlein and Prof. Dr Maaike Endedijk, for their clear guidance and feedback during this study. They sharpened my thinking and taught me how to bring this research to a higher level.

Jointly, I would like to thank all the partners in the BUSLeague project. Their unconditional support and commitment has enabled the development of the tools and knowledge gained with this study.

At last, but not least, I would like to thank my colleagues from the Komma student association for their support and help during this study. I enjoyed working together to improve EST and am proud of all the achievements the association has accomplished. I would like to thank my family and friends for their support during this study. They provided wise counselling and distractions when needed.

Abstract

The transition towards nearly zero-emission buildings necessitates the upskilling of the existing workforce in the construction and energy industry to meet new standards. Micro-training programs are commonly used to update professionals' knowledge, skills, and competencies in this sector. However, the evaluation of training effectiveness and impact is often subjective and conducted by inexperienced evaluators.

This study aims to develop support tools for evaluating training programs in the construction and energy industries. The tools are intended to be versatile, enabling the evaluation of diverse training programs and facilitating the recognition of skills at a European level. A design-based methodology was employed to enhance a rubric for evaluations originally proposed by the University of Twente for the BusLeague project. The tools were co-designed with stakeholders to ensure their viability in real-life contexts. The outcome is a comprehensive toolset and framework that incorporates both objective and subjective measurements. It implements the assessment of Learning Goals and guides the designing and implementation of evaluation programs. The feasibility of the toolset was assessed through iterative developments through piloting and beta-testing.

The results of this study contribute to addressing the need for standardised and reliable methods to evaluate training programs in the construction and energy sectors. By offering a robust framework for measurement and guidance, the toolset holds promise for promoting comparability and recognition of skills across European contexts.

Keywords: upskilling, evaluation, effectiveness assessment, micro-trainings

Table of Contents

1. Introduction

In this chapter, the rationale for the study, the problem statement, social, scientific, and practical implications of the study are presented.

1.1. Rationale for the Study

The transition to more sustainable and energy-efficient practices implies a considerable effort in upskilling the workforce in the construction and energy industries. This is mainly conducted as micro trainings offered by the related organisations.

In the construction industry, evaluations often suffer from a lack of standardisation and structure. Unlike other industries where standardised performance metrics and evaluation frameworks are in place, the construction industry has traditionally relied on ad hoc assessments and informal processes (Jadallah et al., 2021). This means that evaluations are often based on varying criteria, leading to inconsistent and unreliable results. The absence of standardised guidelines makes it challenging to objectively measure and compare performance across projects, teams, or organisations. This lack of standardisation and structure not only hampers the accuracy and fairness of evaluations but also hinders the industry's ability to identify best practices, improve efficiency, and foster meaningful quality control. Efforts to establish standardised evaluation frameworks can help promote transparency and consistency in the construction industry, leading to better decision-making and overall recognition of skills throughout the industry.

The present study will assist organisations in designing structured and scientifically supported evaluations that yield more reliable and comparable results.

1.2. Context of the Study

This study is conducted within the BUSLeague[1] project. BUSLeague is part of the Horizon Europe[2] program, a European Union initiative for research and innovation, aiming to achieve the United Nations' Sustainable Development Goals while also fostering the EU's economic growth. The twelve organisations participating in the consortium represent governmental agencies, universities, formation centres, trade associations and other institutions from seven EU countries. BUSLeague aims at supporting a sustainable cycle of demand and supply for energy efficiency skills in the construction industry. This can be achieved through a series of synergetic initiatives, supporting legislative change, upskilling, and enabling workers in these industries to conduct and deal with new processes, materials, and standards, and stimulating demand for such skills at the market level.

The BUSLeague project succeeds previous Build-Up Skills projects in the European Union's Horizon 2020 program for research and innovation. Preceding projects, such as NEWCOM[3] and BIMplement[4], aimed at mapping the gaps in energy-efficiency-related skills in the construction industry, developing a common qualification platform, and setting up trainings on construction skills. The NEWCOM project developed a framework for standardized units of learning outcomes to facilitate the cross-recognition of skills, whereas BIMplement defined which competences, skills and knowledges were required for a qualified professional in energy transition in the construction industry. In its turn, BUSLeague aims at motivating and upskilling a more comprehensive range of the public, including most of the workforce in the construction and energy sectors, but also the consumer and, ultimately, society.

[1] https://busleague.eu/
[2] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en
[3] https://www.newcomtraining.com/
[4] https://www.bimplement-project.eu/

Upskilling the workforce aims at increasing the competencies in the energy efficiency of the existing workforce in the construction and energy industries. These activities do not only aim at trainings in construction sites but can also emerge in a diversity of activities, such as workshops for employees in construction material stores, online courses for specialists in energy and construction and targeted training in formation centres for professionals in the industry. During the BUSLeague project, several upskilling activities were foreseen to be conducted by the participating organisations. By identifying and analysing the learning outcomes that have been achieved, such as changes in behaviour and practice begotten by those trainings, it should be possible to assess their effectiveness. Hence, the project identified the testing and recognition of skills as one of the main challenges.

As part of the BUSLeague project, this study designed a guidance for the design of evaluations which was pilot-tested and implemented during the project and afterwards. The partnering organisations of the project contributed to the development of this solution by informing the main requirements, opportunities, and boundaries for the implementation of evaluation programs.

1.3. Problem Statement

The endeavour of transitioning to nearly zero-emission buildings presupposes upskilling the existing workforce in the construction and energy industry to meet new standards. In many cases, this is accomplished through short or micro trainings aimed at updating the skills and competencies of professionals. These trainings are offered in different formats, either in digital platforms or face-to-face instruction. To ensure that these learning processes translate into a sustainable transfer of practice, interventions must be evaluated for their effectiveness and impact. Effectiveness can be defined as the extent to which these interventions yield the desired outcomes (Gao et al., 2019), whereas impact relates to long-term and sustained changes in behaviour, practice and results engendered in training (Boulmetis & Dutwin, 2011). However, despite all the advancements in the science of

training (Salas & Cannon-Bowers, 2001), the improvements in the construction industry remained restricted to the assessment of education at the university level and not extended to the assessment of general workforce training programs (Jadallah et al., 2021).

Due to the complexity of the process, most organisations wind up only assessing the effectiveness of interventions and not their impact. As posed by James Kirkpatrick in an interview for the Training Journal[5], organisations usually set a high focus on satisfaction with training rather than on their perceived usefulness. Ultimately, deficient assessment of training programs does not allow for measuring their reliability and validity (Durlak & DuPre, 2008), thereby compromising the efforts invested in training and influencing the outcomes of the learning experience by presenting an inaccurate or biased perspective. Moreover, the absence of standardised methodologies and the divergent criteria used across organisations further impede the comparison and mutual recognition of skills.

The BUSLeague project acknowledged and addressed several of those issues and conducted an exploration of evaluation methods available in literature and those adopted in practice. This identified a significant gap between what is recommended by scientific-based knowledge and the ad hoc practices adopted in the industry, resulting in unreliable and inefficient evaluation methods. During the project, a rubric for an evaluation framework for upskilling in the construction industry was conceptualised to overcome these drawbacks. The rubric suggested the use of Kirkpatrick's four-level model for the evaluation of training programs and compiled a set of validated measuring instruments for the distinct levels. However, the rubric was not being adopted by organisations when designing their evaluations.

This study will seek to understand the reasons hindering the adoption of the evaluation rubric and focus on the development, design, and guidance for the implementation of an evaluation framework for upskilling activities in the construction

5 https://www.trainingjournal.com/articles/interview/means-and-end

industry. To increase its feasibility and usability throughout the industry, the proposed solution needs to be co-designed and tested with the involved stakeholders in multiple iterations. These goals can be achieved through educational design-based research, which presupposes the collaboration of researchers and stakeholders in designing contextualised solutions to address specific problems (McKenney & Reeves, 2018).

Therefore, the following research question was posed:

How to guide the evaluation of the effectiveness of upskilling micro-trainings in energy efficiency for organisations in the construction and energy sectors, with limited (human, financial, material) resources and expertise?

1.4. Practical Relevance

In a recent study, Jadallah et al. (2021) noted that literature on assessment methods in the construction industry is scarce. Their study also pointed out that assessment processes in the construction industry are based on experimental practices, and do not usually account for scientific methods and tools, such as reasonable scales or reliable measurement instruments.

This study will enable organisations in the construction and energy industry to implement practical and feasible scientific-based evaluation processes to assess their training interventions in the energy transition. The results of these evaluations can be used to inform those organisations not only on the outcomes but also on opportunities for improvements in their educational efforts, yielding better training and professional qualification of personnel. Secondly, by establishing and steering organisations towards common practices in evaluations, this study expects to contribute to the facilitation of mutual recognition of skills and qualifications throughout the EU.

The use of design-based research to develop guidance and tools for evaluation programs is expected to contribute to its adoption and applicability by a wider range of organisations, given its productive interactions between researchers and stakeholders. The

findings of this process may inform other future designs for evaluations in this sector. The development of a common scientific-based framework for training evaluations in the construction and energy sectors is also expected to enrich the literature on evaluation processes applicable to the construction and energy industries.

Lastly, this study will not restrict its scope to guiding the design of direct knowledge and skills assessments but also embed elements that allow evaluands to reflect on their learning experiences and knowledge.

2. Theoretical Framework

In this chapter, the main conceptual underpinnings and key constructs that encompass the study will be presented, drawing upon established theories and models in the field.

2.1. Program Evaluation

Across literature, program evaluation is generally described as a rigorous, systematic investigation of the effectiveness or impact of an educational or training program, reliant on scientific methodology (Shek et al., 2018). McBride (2018) notes that this investigation is not only a process, but it can also be a discipline and an intervention in itself. The outcomes of evaluations also differ in their aimed purpose. From a process perspective, evaluations determine whether programs achieved their primarily stated goals. Conversely, from a research perspective, they judge and assess the quality of programs, or collect data that will inform decisions made by stakeholders (Chen, 2018).

Larson and Berliner (1983) describe evaluations as a system composed of three main elements: inputs, processes, and outcomes. Inputs relate to all the resources and methodologies invested in the evaluation process, which can be paralleled to design requirements. Evaluation processes are closely related to all elements in the implementation phase of an evaluation, such as the interaction between evaluators and evaluands and the data collection process. The authors define the outcomes of evaluation as the impact they produce, represented by the decisions that were influenced by the results of the process.

The effectiveness of upskilling interventions in conventional learning can be defined as the degree to which desired outcomes are achieved (Ho & Dzeng, 2010; Gao et al., 2019). Hattie (2008) emphasises the significance of having specific goals, intentions, and criteria for success when considering the value of self-assessments and evaluations. Furthermore, Hattie points out the compelling evidence that supports the positive impact of goals on enhancing performance and self-efficacy.

Different approaches are used when designing evaluations, using quantitative, qualitative, or integrated methods to add precision and depth to the results. One of the most consolidated program evaluation frameworks used by the industry is the four-level Kirkpatrick model, which has been widely accepted since its proposal in the late fifties due to its simplicity.

2.2. Four-Level Evaluation Model

Kirkpatrick's (2006) Four-Level Evaluation is a well-known and established model for assessing the effectiveness of training programs. The model consists of measurements at four consecutive levels: reaction, learning outcomes, behavioural change and impact or results. At level 1 – reaction – the satisfaction of participants with the training received is measured. It does not measure the amount of acquired skills or knowledge gained by learners, but rather their motivation, interest and attention levels (Smidt et al., 2009). Kirkpatrick (2008, p.22) recognises that the level of satisfaction may not be a reliable predictor of effectiveness, but it is a known antecedent to it. Level 2 measures the learning outcomes in the form of objective knowledge or skill tests applied to evaluands, whose results may be used to infer the effectiveness of a training program. Level 3 measures the behavioural change resulting from the training, through the observation of long-lasting changes in attitudes and behaviours. These changes, which imply the use of the acquired knowledge and skills in new contexts and situations, are representative of the transfer of learning (Steiner, 2001). The fourth level measures the impact caused by an intervention in the operational and financial results of an organisation, using analysis of productivity, quality increase or revenues. Thus, the evaluation for the first three levels (satisfaction, learning and behaviour) must be designed jointly and must be consonant with the training intervention, whereas level 4 (results) is typically measured using external and existing indicators and does not require designing or implementation (see Figure 1).

Figure 1

Kirkpatrick's 4-Level Model

	Kirkpatrick's 4-level model	**Measurement**
Level 1	Reaction	Satisfaction, engagement, relevance.
Level 2	Learning	Knowledge and skills
Level 3	Behaviour	Behavioural change, transfer of learning
Level 4	Results	Impact, tangible results

Since its first proposal, over six decades ago, the Four-Level Evaluation model has been subject to the scrutiny of research. In an article reviewing literature on the model, Alliger and Janak (1989) identified and discussed three major assumptions, that could lead to misleading generalisations. The first one refers to the hierarchical nature of the model. As each level subsumes the previous levels, it could be assumed that the top tiers are more informative and thus, leading to performing analysis only at higher levels and generalising the effectiveness of training programs based on their perceived impact. However, Kirkpatrick and Kirkpatrick (2006, 2008) explicitly denote that the model should be followed from bottom-to-top, and that the sequence levels one to four should be respected.

The second assumption is that each level is caused by the previous level, but causality is usually hard to either systematically prove or disprove. The third assumption is that there is always a positive correlation between a level and the precedent one, i.e., the results of the evaluation at one level could be predictors of the results of the subsequent levels. Thus, the second (causality) and third (positive intercorrelation) assumptions are related. The model assumes that all levels are causally connected and positively intercorrelated. This would imply that positive reactions would translate into positive learning outcomes, leading to behavioural change and resulting in positive impacts on organisations. Other critiques point out that the Four-Level Evaluation model may not be sufficient to

respond to current evaluation needs and does not properly answer the questions organisations have about the effectiveness of their training programs (E. A. Ruona et al., 2002). Conversely, the model is still widely used throughout organisations. This popularity might be attributed to its relatively simple taxonomy and practicality (Alliger et al., 1997; Frye & Hemmer, 2012; Smidt et al., 2009).

Lastly, in his critical analysis of the four-level model, Bates (Bates, 2004) noted that the model does not account for contextual factors, such as characteristics of the organisation and training design and delivery methods, implying that the model does not consider them essential for effective evaluations. It also must be noted that the Kirkpatrick model does not consider self-reported measures at all levels, and literature on their use in the model is scarce.

2.3. Self-Reported Measures of Performance

Self-reported measures relate to the two concepts of self-evaluation and -assessment. Self-evaluation can be defined as the individual's subjective judgement or appraisal of their work, usually against a known or defined set of criteria (Rolheiser & Ross, 2001), whereas self-assessment depicts the ongoing introspective practice of self-reflection that prompts individuals to their behaviours, modulating and regulating their experiences (Pisklakov, 2014). They generally contrast with objective measures in assessments, which refer to quantitative or measurable criteria used to evaluate performance, knowledge, or skills in a standardised manner. However, this terminology is not consensual throughout literature, and different authors attribute different definitions or nuances according to their focus of studies (see Dauenbeimer, 2002; Rohlheiser & Ross, 2001, Sedikides, 1993). This study will adopt Rohlheiser's (1996) understanding of self-evaluation, which encompasses self-assessments.

The positive effects of both self-reported measures seem to be consensual throughout literature. Studies have found that self-evaluations potentially impact learners'

performance, self-efficacy, and intrinsic motivation (Rolheiser & Ross, 2001). Self-assessment has been found to increase the certainty of self-knowledge (Sedikides & Strube, 1997) and of presenting itself as a learning opportunity (Pisklakov, 2014). Andrade (2019) has pointed out that literature sustains a positive association between self-assessment and learning outcomes. Rolheiser (1996) posits that self-evaluation affects self-confidence. A positive impact on self-confidence may lead learners' to set higher goals and commit more effort to learn, which may yield higher achievement. The same phenomenon was noted by Pisklakov (2014) in a study on the role of self-evaluation and self-assessment in medical education.

The reliability of self-evaluations has long been debated, with inconclusive results. In a meta-analysis on the validity of self-reported scores, Kuncel et al. (2005) revealed that the reliability of the scores increased at higher educational levels, implying that more advanced or educated learners dispose of more tools to perform such assessments. The authors also point out that studies have shown that self-scored grades are frequently consistent predictors of actual grades, which implies that the validity of self-evaluations cannot be excluded. Self-reported data obtained from student experience surveys have been shown to have a good correlation with student Grade Point Averages (GPAs) in the US, and they perform better than standardised tests, as demonstrated by empirical evidence (Arico et al., 2018). Conversely, it has been argued that individuals may also be overconfident about their performance in learning, and attribute higher self-scores (Anaya, 1999; Rogaten & Rienties, 2021; van Uum & Pepin, 2022). As noted by Schunk (2010), self-evaluations of progress have the potential to increase individuals' sense of self-efficacy and motivation. However, adequate self-evaluations are only obtained when individuals compare their performance to the learning goals before the learning experience.

2.4. Learning Goals

Despite sometimes being used interchangeably (Marzano, 2009), the terms learning goals and learning objectives have adopted distinct conceptualisations throughout literature. In general, learning goals are used to refer to the broader, general possible intended outcomes of a learning activity (Fessl et al., 2021; Marzano, 2009; Mcnall, 2018), whereas learning objectives describe more specific, observable or tangible outcomes, expressed in behaviours resulting from a learning experience (Adams, 2015a; McNall, 2018; Schiekirka et al., 2013). Concurrently, it has also been noted that different stakeholders may have distinct, and sometimes divergent understandings of the learning goals and objectives of an educational or training program (McNall, 2018). In this study, the term Learning Goals will be used to describe the intended and observable learning outcomes resulting from a training.

In his book – Visible Learning - synthesising over eight hundred meta-analyses related to achievement, Hattie (2008) states that having clear goals, intentions and criteria for success is a claimed basis for the value of self-assessments and evaluations. He also points out that there is compelling evidence of the impact of goals for enhancing performance and self-efficacy (p.163 -165). However, the formulation of learning goals and objectives and their translation into clear, effective statements requires effort and attention from designers of instruction and instructors (De Long et al., 2005). In a study on the impact of explicating learning goals, Fessl (2021) found that about 50% of goals set by university teachers were poorly formulated, i.e., they contained redundant or unnecessary content for the intended final learning outcomes.

Studies that prescribe a systematic for the formulation of learning goals (Chatterjee & Corral, 2017; Fessl et al., 2021; Marzano, 2009) generally agree that effective learning goals consist of a clear statement, declaring what is expected to be performed by an individual at under specific circumstances, such as after a training intervention. Most studies also adopt a taxonomy (Ferguson, 1998; Fessl et al., 2021; Marzano, 2009; Wei et al., 2021) for the

classification of the level of learning and selection of one measurable item. Bloom's Revised Taxonomy (Anderson & Krathwol, 2001) is the most widespread tool used, either in its full format or in adaptations derived from it (Fessl et al., 2021; Marzano, 2009).

The SMART criteria is a goal-setting framework developed by George T. Doran in the 1980s (Doran, 1981). SMART is an acronym for *Specific, Measurable, Achievable, Relevant*, and *Time-bound*. This framework provides a systematic approach to setting effective goals, which are clear and well-defined. Measurable goals have tangible criteria to track progress. Achievable goals are realistic, and within reach, relevant goals are aligned with broader objectives, and time-bound goals have a clear deadline. The SMART criteria help to define meaningful goals that are focused, trackable, attainable, relevant, and time-bound, enhancing their chances of success.

John Biggs (J. Biggs, 1999), who developed the constructive alignment theory, defended that teaching and learning should be aligned with the assessment to create a coherent and effective learning experience. This alignment is only enabled by the clear definition of the intended learning goals, which are to guide the design of teaching methods, materials, activities, and assessments.

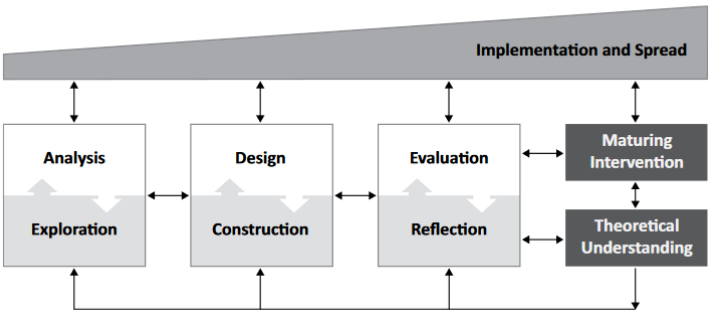
3. The Present Study

This study will use educational design-based research (DBR) (McKenney & Reeves, 2018) as its primary research approach. DBR is well-suited for studying complex educational interventions or innovations, and it emphasises collaboration between researchers and practitioners to design, implement, and evaluate educational interventions in real-world settings. Through an iterative process of design, enactment, analysis, and refinement, DBR seeks to create effective, sustainable, and scalable educational innovations that can be widely adopted and replicated in diverse contexts.

DBR will be used to design, implement, and evaluate an intervention aimed at improving student engagement and learning outcomes in a particular subject area, to produce practical knowledge that can inform teaching and learning practices in similar contexts. McKenney and Reeves (2018) proposed a generic model for educational design research, which aimed at bridging the gap between the knowledge acquired in educational science and its applicability in real-life contexts, whilst increasing theoretical understanding and promoting innovation. Their model consists of three core phases (analysis and exploration, design and construction, and evaluation and reflection), plus a phase of maturing intervention and theoretical understanding that take place during later phases of information and spread (see Figure 2).

Figure 2

Generic Model for DBR (McKenney & Reeves, 2018).



The model prescribes that the structure of DBR is flexible and iterative. Each phase provides information and is fed by the others. After the completion of a cycle, a ponderation of the processes and results is carried out, informing the next iteration.

The analysis and exploration phase is characterised by the definition of the practical problem. The main activity of the analysis is to refine the causes of the problem through the analysis of its context, stakeholders, and the identification of their real needs. During this phase, literature research is conducted to determine which aspects related to the problem have already been studied or identified by scientific knowledge. Concurrently, in exploration, it is sought to understand how similar challenges have been dealt with and look up for possible solutions.

The design and construction phase aims at mapping and exploring possible solutions, concerning the main requirements and propositions. Design engages in the process of generating ideas and exploring potential solutions, whereas during construction the prototypes are built and tested. This phase may lead to two possible outputs: documents prescribing potential designs, with their specifications; or constructed solutions, such as tools, materials, interventions, or guidelines.

Finally, at the evaluation and reflection phase, the artefact is systematically assessed for its theoretical robustness, or for its viability by try-outs it in context. The evaluation is usually focused on the processes or outcomes engendered by the artefact and is aimed at improving or verifying its quality. The reflection process aids in better understanding the intervention, the appropriateness of its intentions, and the effects it produces.

This study aimed to design guidance for the development and implementation of evaluations for training activities related to energy efficiency in the construction industry, as part of the objectives of the BUSLeague project. Each phase of the study will be presented separately, providing detailed explanations of the methods, participants, instruments, materials, analysis, and results utilised. Furthermore, each phase will be guided by

secondary research questions that were formulated to assist in answering the main research question. An overview of the phases and activities can be seen in Table 1.

Table 1

Overview of the phases of this study.

Phase 1 **Analysis and Exploration** Two pilots	**Phase 2** **Design and Construction** Two pilots Two workshops	**Phase 3** **Evaluation and Reflection** One workshop
RQ1: a. What are the current hindrances in the evaluation rubric, and b. What are the capabilities, limitations and needs of stakeholders in designing evaluation programs? **Focus:** Identifying and defining the causes of the problem (designing evaluation programs following the rubric) and identifying possible solutions.	**RQ2:** Which structure and tools can support the autonomous development of evaluation programs for organisations in the construction industry? **Focus**: Designing tools following design requirements, usability, and feasibility.	**RQ 3**: How do these designs perform in guiding the design of evaluation programs in organisations in the construction and energy efficiency sector in a beta test? **Focus**: Assessing the effectiveness of design – Beta testing.
Analysis	**Design**	**Evaluation**
Methods: Document analysis **Products:** Analysis and interpretation of existing data, the definition of requirements and propositions.	**Methods:** Generation, consideration and checking of ideas. Designing prototypes. **Products:** Creation of a skeleton design, creation of detailed specifications.	**Methods**: Beta-test of collection of tools and guidance (*one workshop*) with all partners. Survey, Plus and Minus test. **Products**: Assessment of usability and perceived usefulness of developed tools.
Exploration	**Construction**	**Reflection**
Methods: Piloting (*two pilots*). **Products:** Problem definition, long-range goal, partial design requirements, initial design propositions.	**Methods:** Piloting and trialling workshops (*two pilots* and *two workshops*). **Products:** Initial trialling of prototypes of tools.	**Methods:** Structured reflection on the DBR process and outcomes. **Products:** Report and discussion.

In *Analysis and Exploration* (phase 1), the objective is to gain a deeper understanding of the problem. The chapter begins by presenting relevant questions, which are then explored through document analysis, literature research, and piloting an existing rubric. The outcomes of these activities include problem definition, a long-term goal, partial design requirements, and preliminary design propositions. The analysis and exploration phase involved studying the project's scope, objectives, and requirements through document analysis and literature research. Two pilots, consisting of semi-structured interviews, were conducted to understand the obstacles in using the rubric and the challenges in evaluation design. The activities aimed to identify the shortcomings of the rubric and the organisations' capabilities in designing evaluation programs, while literature research provided insights and support for decision-making in developing a solution.

The focus in *Design and Construction* (phase 2) is on the co-designing and construction of the framework and guidance tools for the design of evaluations. The objective of this phase is to develop an initial prototype of an evaluation framework and associated tools. To achieve this, various activities are implemented to explore potential solutions to the problem, such as generating, considering, and verifying ideas. The methods used to achieve this goal include brainstorming sessions, creating strengths/weaknesses matrices, and logic modelling. The proposed solutions are then mapped into a design by defining requirements and propositions, creating a skeleton design, and developing detailed specifications. Overall, this phase involves a structured and comprehensive approach to the development of an evaluation framework, which is informed by rigorous research and design methodologies.

In *Evaluation and Reflection* (phase 3), the assembly of all designed tools is presented and tested with the stakeholders at a workshop. The goal of this phase is to evaluate the usability and feasibility of the proposed tools in guiding the design of evaluations for upskilling activities in the construction industry. Participants are asked to use the tools and respond to a survey. The artefacts produced during the workshop are also assessed to determine the quality of the results obtained with the assistance of the tools.

This phase concludes with a reflection on the results achieved and recommendations for future studies or developments.

3.1. Participants

Interviewees in this study corresponded to the individuals assigned as head representatives of the corresponding organisations for trainings and evaluations that occurred during the length of the project. The pool of twelve partners is constituted of organisations directly related to the energy efficiency transition sector, such as educational institutions, regulatory agencies, and construction-material stores.

The collection and treatment of data for the study were conducted under the ethical approval of the Faculty of Behavioural, Management and Social Sciences of the University of Twente, n.220198.

4. Analysis and Exploration

The goal of this phase was to determine the type of guidance required to assist organisations in the construction industry in implementing upskilling evaluations. To determine adequate guidance, this study needed to define the causes hindering the adoption of the rubric and identify possible solutions. To achieve this, two approaches were adopted: (1) a document analysis from the documentation of BUSLeague, and (2) the piloting of the design of evaluations. The outcomes of these activities were synthesised and resulted in four main products: the definition of the problem, the long-range goals, the partial design requirements, and the initial design propositions.

To understand the discrepancy between the current and desired situations on the problem of guiding the evaluation in the construction industry, the following research questions were formulated:

RQ.1a: *What are the current hindrances in the evaluation rubric?*

RQ.1b: *What are the capabilities, limitations and needs of stakeholders in designing evaluation programs?*

4.1. Document Analysis

A document analysis was performed to answer the proposed research questions. Selected documents comprised reports and other publications from the BUS League project. This qualitative data aimed at deepening the understanding of the current situation and refining the problem according to the participating organisations.

4.1.1. Methods of Document Analysis

The document analysis procedure started with the selection of relevant documents. To determine which documents were suitable to answer the research questions, a set of criteria was established. To respond to RQ.1a, documents should mention the *rubric for evaluation*. Additionally, to respond to RQ.1b, documents should report on *upskilling*

activities (*e.g., trainings, learning objectives, outcomes, etc.*) and *evaluations* (*e.g., evaluation program, assessments of skills, knowledge, and competencies, etc.*).

The selection criteria were applied to a total of eighteen documents made available by the project. After careful analysis, three documents were considered relevant to the study and were further analysed:

1. Report Task 5.3 Measure Effectiveness and Timeliness from Educational Perspective
2. Report D2.6 Defining Personal Recognition for each Country.
3. Report: D27 (D5.3): Report on BUSLeague Activities from an Anthropological Perspective.

4.1.2. Data Analysis of Documents

To generate qualitative data, this study Mayring's (2014) qualitative content analysis method, as it offers a systematic and structured approach for examining the content of various textual materials, including interviews, documents, and open-ended surveys. An inductive coding scheme was created for the analysis of the documents. Based on the research questions, four main themes were defined: (1) hindrances in the use of the rubric, (2) the capabilities, (3) limitations and (4) needs of the organisations in terms of the design of evaluations. Each theme comprised several sub-themes identified in literature (Durlak & DuPre, 2008; Jadallah et al., 2021; Kirkpatrick & Kirkpatrick, 2008; Reeve & Peerbhoy, 2007) as seen in Table 2, and were defined as coding rules (see full description in Appendix A).

Table 2

Main Themes and Subject Items of Document Analysis

Main Theme	Sub-themes
RQ.1a - What are the current hindrances in the evaluation rubric?	
Hindrances	Lack of Awareness
	Resistance to Change
	Complexity and Difficulty
	Resource Constraints
	Inadequate Training and Support
	Compatibility Issues
	Lack of Stakeholder Engagement
	Perceived Irrelevance or Ineffectiveness
	Legal and Regulatory Barriers
	Other
RQ.1b - What are the capabilities, limitations and needs of stakeholders in designing evaluation programs?	
Capabilities	Expertise in Evaluation Design
	Skilled Evaluation Team
	Resources
	Data Collection and Analysis
	Adaptability and Flexibility
	Planning and Implementation
	Other
Limitations	Expertise in Evaluation Design
	Skilled Evaluation Team
	Resources
	Data Collection and Analysis
	Adaptability and Flexibility
	Planning and Implementation
	Other
Needs	Training and Capacity Building
	Access to Expertise
	Clear Guidelines and Frameworks
	Collaboration and Knowledge Sharing
	Adequate Resources
	Data Collection and Analysis

4.1.3. Results

In the following sections, the results from the analysis of the three selected documents will be presented and interpreted.

4.1.3.1. Results of the Document Analysis: Task 5.3 Measure Effectiveness and Timeliness from Educational Perspective.

The rubric for evaluations of upskilling in the construction industry was established at earlier stages of the BUSLeague project and was based on a simplification and adaptation of the Kirkpatrick model for training evaluation.

Before identifying the hindrances in the use of the rubric and the capabilities, limitations and needs of organisations in designing evaluations, the document analysis sought to characterise the training activities the participating organisations conducted. The documentation of the project revealed that most interventions consisted of micro trainings, or short modules aimed at blue-collar workers, and were planned to take place either online or in face-to-face instruction (Figure 3).

Figure *3*

Characteristics of Interventions Per Country

Characteristics of interventions per country											
	BG			NL					IE		SP
	1. NZEB Training	2. Individual upskilling courses at construction fairs	3. Online training for homeowners' associations	1. EE-Skills	2. Girlsday events	3. Upskilling in DIY-stores on awareness	4. Cooperation with Breman academy and Breman Breinn	5. Promotion of the BUS-app	6. NZEB trainings	7. Retrofit skills	8. Bauhaus
Target group = workers	x	x		x		x	x	x	x	x	x
Currently running			x					x	x	x	
Starting within 1 year	x	x		x		(x)	x				X
Training/course for professionalization	x	x	x	(x)		x	(x)		x	(x)	x
Information available in Moodle	x			x					x		x
Information available in BUS-app				x			x			x	
User review	x			x	x		x		x	x	x

Note. Information is incomplete when a cross is in brackets (x).

From: Evaluation framework of upskilling in the construction sector V0.5 (2021).

The rubric focused on the three levels of Kirkpatrick's model where evaluations must be designed and require support (1-Reaction, 2-Learning, 3-Behavior). The levels match

Kirkpatrick's model and were named Level 1 – Reaction, Level 2 – Learning Outcomes; and Level 3 – Transfer of Learning (see Figure 4). The focus of levels 1 and 2 is on measuring the effectiveness of trainings, and level 3 measures their impact. In compliance with the objectives of the BUSLeague project, which also aims for the facilitation and recognition of energy-efficiency-related skills throughout Europe, the rubric also suggests the implementation of both subjective and objective measurements for assessments, in line with the recommendations of the European Qualifications Framework[6]. These measurements not only inform on the effectiveness and impact but also help identify opportunities to improve the design of the intervention and the supporting context for successful transfer. Secondarily, subjective measurements, such as self-reported scores, are easy to implement and can be used as indicators whenever a more precise measurement is not available (Andrade, 2019).

Figure 4

Comparison Between Kirkpatrick's Model and the Rubric for Evaluation

	Kirkpatrick's 4-level model	Measurement	Rubric for evaluation
Level 1	**Reaction**	Satisfaction, engagement, relevance.	**Satisfaction**
Level 2	**Learning**	Knowledge and skills	**Learning Outcomes**
Level 3	**Behaviour**	Behavioural change, transfer of learning	**Transfer of Learning**
Level 4	**Results**	Impact, tangible results	

Whereas Kirkpatrick's model is flexible and largely adopts experimental survey tools and assessment techniques, the proposed rubric adopted scientifically tested instruments for measuring reaction and design principles for training programs. For Level 1, a series of instruments that indicate the level of satisfaction of learners both in terms of perceived utility and relevance was selected. At level 2 (learning outcomes), the use of Bloom's revised

[6] https://europa.eu/europass/en/europass-tools/european-qualifications-framework

taxonomy (Oliver et al., 2004) is recommended for guiding the assessment of learning outcomes. At Level 3 (change in behaviour due to training), the addition of a self-perception survey to the observational assessment is suggested, based on the work of Chauhan et al. (2016). The rubric also comprised a tool for evaluating design principles, based on the study of Saks and Belcourt (2006) that investigated and identified the most influential variables for transfer of training in organisations.

However, due to their different levels of expertise, the conceptualisation of a framework was deemed insufficient to respond to the practical needs of organisations in evaluating the effectiveness of training programs. According to a survey carried out upon the first introduction to the rubric, it was considered valuable and insightful. The rubric was praised for containing both evaluand-related items and tools for assessing design principles, which indicated the robustness of a training program. Conversely, the implementation of the rubric was considered at times too complex for the inexperienced evaluator.

Numerous challenges were identified. It was stated that organisations in the construction industry do not usually have continuing professional development programs, especially for blue-collar workers, and are not motivated to invest in training for those professionals. Training is deemed as costly, and it is only adopted when there is a real need for learning. Organisations also reported that blue-collar workers tend to have lower motivation for learning, as they do not perceive the need for upskilling. It was also noted that most blue-collar workers are unfamiliar with digital learning environments. A documented interview research carried out within the report also revealed that organisations were uncertain of the effectiveness of upskilling activities and how they should be measured. Secondly, the interviews also provided a general view of what they considered as requirements for upskilling interventions. Most organisations replied that they thought trainings should be engaging, brief, attractive, straightforward, user friendly and that learning gains should be visible to users and employers (see Appendix B).

When asked to clarify their plans for training, most organisations were unclear about their training goals. Some enumerated the intention to perform several trainings, but only a few were clearly defined had their goals stated. The audience for those trainings is also very diverse, ranging from highly skilled white-collar professionals to craftsmen. Notably, the interviews revealed that organisations consider that learning gains should be visible to both providers of training and trainees, and that all interviewees declared using electronic spreadsheets as their main tool for data management and analysis.

To address the complexity and facilitate the adoption of the rubric, a solution proposed is to co-design a framework in collaboration with partnering organisations. In the context of evaluation, a framework refers to a structured and systematic approach that provides a set of elements or requirements for conducting an evaluation. It outlines the essential components and considerations necessary to evaluate a program, project, or intervention effectively. A framework serves as a guide or structure to ensure that the evaluation process is comprehensive, consistent, and aligned with the evaluation objectives.

By involving partnering organisations in the design process, their expertise and input can be leveraged to ensure the framework aligns with the specific needs and goals of different organisations, ultimately leading to more meaningful and useful evaluations of training programs. This approach would allow for a more comprehensive understanding of the practical needs and challenges faced by organisations in evaluating their training programs, resulting in a more effective and tailored framework.

4.1.3.2. Results of the Document Analysis: Report D2.6 Defining Personal Recognition for each Country.

As a reference document, this report sets the criteria for personal and mutual recognition of skills at the EU level for organisations in the construction industry. This recognition of skills can only be obtained through adequate and commonly agreed standards for evaluations. The document implies the adoption of the standards of the NEWCOM project, which preceded the BUSLeague project, and sets a methodology based

on the definition of Units of Learning Outcomes (ULOs). These units are equivalent to learning goals, and their structure is based on Bloom's revised taxonomy (Anderson & Krathwol, 2001). Thus, it also sets the minimum requirements in terms of the capabilities of organisations in conducting evaluations, such as the structuring of upskilling programs based on learning goals, adequate evaluations, and the ability to adapt their design to fit specific contexts. Overall, this document clarifies what requirements organisations should meet when designing upskilling activities and evaluation programs to meet the standards defined by the project.

4.1.3.3. Results of the Document Analysis: Report: D27 (D5.3): Report on BUSLeague Activities from an Anthropological Perspective.

This report reflected from an anthropological perspective, the impact of a selection of upskilling interventions carried out in the project. It provided insights on a few limitations organisations faced during the implemented interventions, most notably the lack of motivation, either financial or derived from knowledge of the benefits of upskilling activities and recognition of skills. However, despite mentioning that evaluations were performed, the report does not specify the type of assessments carried out, and results are generally described as 'satisfactory' or 'positive'. The context in which they are used, however, could imply that some evaluations were based on subjective measurements.

4.1.4 Synthesis of Document Analysis

A synthesis of the knowledge and insights gained from the document analysis can be found in table 3.

Table 3

Synthesis of Document Analysis

RQ.1a - Hindrances in the Evaluation Rubric

Hinderances	Lack of Awareness	Organizations in the construction industry may lack awareness of the benefits and importance of upskilling activities, which hinders their motivation to invest in training programs.
	Resistance to Change	Organizations exhibit resistance to change, making it challenging to adopt new approaches to training and evaluation.
	Complexity and Difficulty	The implementation of evaluation rubrics and frameworks is perceived as complex, especially by inexperienced evaluators, which hampers their adoption and effectiveness.
	Resource Constraints	Organizations face resource constraints, including limited financial resources, time constraints, or a lack of necessary tools and materials for conducting evaluations, impacting the quality and implementation of evaluations.
	Inadequate Training and Support	Organizations lack training and support in evaluating training programs, which affects their ability to design and conduct evaluations effectively.
	Compatibility Issues	Blue-collar workers face challenges in adapting to digital learning environments, posing compatibility issues and hindering the implementation and effectiveness of upskilling programs.
	Lack of Stakeholder Engagement	Insufficient involvement of key stakeholders in the design and evaluation processes hampers the successful implementation and acceptance of upskilling programs.
	Perceived Irrelevance or Ineffectiveness	Some blue-collar workers do not perceive the need for upskilling or the relevance of training programs, affecting their motivation to participate.

RQ.1b - Capabilities, Limitations, and Needs of Stakeholders in Designing Evaluation Programs

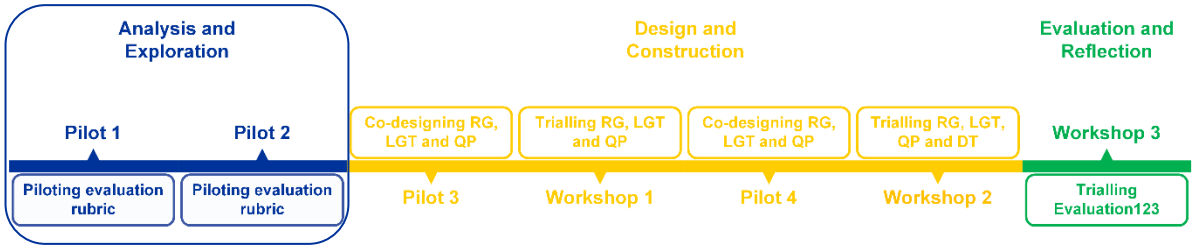
Capabilities	Resources	Adequate resources are required to carry out evaluations effectively.
	Data Collection and Analysis	Organizations can only collect and analyse descriptive data
	Adaptability and Flexibility	Organizations should be able to adapt the design of evaluation programs to fit specific contexts.
Limitations	Expertise in Evaluation Design	Organizations may face limitations in terms of their expertise in designing evaluation frameworks and rubrics.
	Skilled Evaluation Team	Limited availability of skilled personnel in conducting evaluations can be a constraint.
	Data Collection and Analysis	Limited capacity in data collection and analysis can be a limitation.
	Adaptability and Flexibility	Lack of flexibility and adaptability in evaluation program design can pose limitations.

4.2. Piloting of Design of Evaluations

To unveil the practical issues involved in the design of evaluations, this study conducted piloting sessions as field-exploratory activities (McKenney & Reeves, 2018). Two pilot sessions were conducted to understand needs and problems organisations were facing when designing their training evaluations with the rubric. This included understanding the shortcomings of the rubric as well as the capabilities, limitations and needs of organisations in terms of evaluation design. This information was needed to complement the results of the document analysis and allow the proper definition of the problem, its context and for defining possible solutions. An overview of the pilots and workshops carried out in this study can be seen in Figure 5.

Figure 5

Overview of Pilots and Workshops Carried Out In This Study.



4.2.1. Participants in Piloting of Design of Evaluations

Pilots were conducted for two interventions. In the first pilot, carried out in Spain, one evaluator from a construction material store and one evaluator from a public-interest organisation partnered in designing an evaluation for a module consisting of 10 micro-lectures. Pilot two was carried out in Austria with an evaluator from an energy agency for a micro training. These evaluators were the professionals responsible for managing the training activities for their organisations.

4.2.2. Methods of Piloting Design of Evaluations

During the pilots, evaluators were assisted in the application of the rubric for the design of evaluations. The collection of data was performed through semi-structured interviews at the beginning (briefing) and the end (debriefing) of each pilot. A specific guide was developed for the interviews in *Analysis and Exploration* (see Appendix C). The briefing component aimed at gaining knowledge on the type and content of trainings provided by organisations in the construction industry. This first part included information such as the developer of the training, the target audience, type of training, and the learning elements contained in the program. To understand the context of workplace learning in organisations, training programs were assessed using a rubric inspired on the work of Saks and Belcourt (2006) on training activities and transfer of training in organisations. It also included eight open-ended questions, such as: *What needs to be evaluated? What would the ideal approach be? What type of information is expected to be obtained? What kind of decisions are influenced by such data?* This sought to determine the wishes, needs and limitations of the stakeholders regarding their training programs.

The second part of the interviews occurred at the debriefing sessions and consisted of thirty-eight questions related to different aspects of the evaluation process. Its focus was on unveiling what type of guidance and tools the organisations deemed needed for them to implement the Evaluation Framework for Upskilling in the Construction industry autonomously. Questions were clustered into five segments: initial expectations, results, evaluation process, support, and rubric. The first segment - initial expectations – tried to elucidate the expectations towards the evaluation. Questions in this segment included *what was your initial desire/intention for this evaluation process?* The results segment aimed at determining the capability of participants in interpreting the results obtained and their perceived usefulness through questions such as: *what did you learn from the data?* The following segment, evaluation process, focused on understanding the constraints and difficulties perceived by participants in designing and implementing an evaluation program.

Participants were asked, e.g., *Did you experience any difficulties in the process*? *and Is it feasible to adopt this type of assessment in a realistic context?* In the support segment, it was tried to define where and what type of support was most required. This was clarified in questions such as: *If you were to implement this by yourself in the future, what kind of support would be needed*? Lastly, a segment named rubric aimed at assessing the use of the rubric proposed by the Evaluation Framework for Upskilling in the Construction industry, by asking participants, for example, *do you think the rubric suited well the purpose of the evaluation?*

The answers provided to the semi-structured interviews were later coded using the same scheme developed for the document analysis. This allowed us to fill the gaps in knowledge not covered by the documentation and refine the understanding of the needs, limitations, and capabilities of organisations in designing evaluation programs, as well as identifying the factors preventing them from adopting the use of the rubric.

4.2.3. Integrated Results from Pilots 1 and 2

Pilots 1 and 2 sought to establish an understanding on the hindrances, and opportunities present during the design of evaluation programs informed by the rubric. The key findings of the pilots, including the lessons learned from them and the implications are here summarised. During the two pilots, the following challenges and themes have been identified:

- **Misalignment of evaluation goals and capabilities.** In both pilots, the evaluators' initial goal statements were to determine the effectiveness of the training in enabling participants to perform new tasks based on the acquired skill and knowledge. However, these goals imply an evaluation at level 3 (transfer of learning) and can only be done at least a few weeks after the upskilling activity has taken place. Assessing skills considerably demands more resources (human, material, and time) than the ones afforded by the evaluator.

- **Formulation of measurable learning goals.** The learning goals for the training activity were not previously stated and had to be clearly defined. The definition of learning goals was considered time-demanding and complex by the respondent. Unguided attempts to define learning goals usually led to overly broad statements, with multiple and distinct measurable elements, which precludes reliable assessments.
- **Understanding of the evaluation rubric.** Upon the presentation of the results, in the debriefing session, the feedback of the evaluation process was positive. When asked about the evaluation process and its feasibility, respondents noted that the process is rather complex and requires facilitation. The rubric, depicting the three levels and selected instruments is essential for assuring reliable and valid measurements, but its structure was deemed insufficient for its autonomous implementation. The amplitude of rubric, which typically contains multiple items for measuring each construct, was perceived as redundant, as evaluators would prefer quick and concise assessments. Participants also did not initially understand the value of self-reported assessments of knowledge, skills, or competences.
- **Analysis of results**. It was stated that even highly educated professionals in the field are not familiar with statistical tools, and it was not expected that other professionals in related areas would be able to run statistical analysis. According to the interviewee, data in the sector is mostly managed in spreadsheets, and presented and analysed descriptively. This loss in accuracy is considered acceptable, in view of the gains in simplicity and agility when interpreting data.

4.3. Synthesis of the Analysis and Exploration Phase

The processes carried out during Analysis and Exploration created a deeper understanding of the problem in guiding the design of evaluations both from theoretical and practical perspectives. As a result, the following four outcomes have been achieved:

4.3.1. Problem Definition

Surveys conducted with the organizations at earlier stages of the project revealed that their intentions for upskilling activities were often still vague and incomplete. Most partners were only able to declare a few elements of a program, such as content, type of training audience, structure, scheduling, or evaluation proposals. The reported upcoming upskilling activities aimed mostly at blue-collar workers and other professionals. The diversity of types of organisations participating in the project (e.g., construction material stores, professionalisation training centres, regulatory agencies, etc.) with different levels of expertise must be taken into consideration when proposing a solution.

Pilots have revealed that organisations do not follow a defined or structured methodology for designing their evaluations, relying in replicating experimental models or developing a new one when needed. The importance of following systematic and methodological processes for evaluating is consensual in literature (Wanzer, 2021). Scriven (1994) alerts that arbitrarily designed evaluations are misleading and may lead to consequential and costly decisions.

The Evaluation framework proposed in the BUSLeague project had not yet been adopted because it was deemed overly complex, extensive, and lacked specific tools for its implementation. To be able to use it, organisations needed procedural instructions, guidance, and easy-to-follow steps. Secondarily, it was noted that the learning goals for training activities are not always clearly defined or stated. Learning goals are central to the assessment or learning outcomes (Marzano, 2009), and their lack of definition affects the ability of organisations in reliably measuring the results of their training efforts.

Therefore, this educational design research study aims at providing guidance for organisations in the construction industry to design evaluations using appropriate methods and instruments. Their main constraints relate to the lack of fundamental knowledge on evaluation programs and adequate tools to assist them in those processes. Thus, the initial

identified problem of understanding the hinderances in the adoption of the rubric and providing guidance for the design of evaluation programs has been specified. Thus, this study will focus on the development of an evaluation framework and design of tools that will support organisations in designing evaluations for upskilling activities.

4.3.2. Long-Range Goals

As a result, the long-range goal for this study is to develop and implement a comprehensive evaluation framework and tools for upskilling programs in the construction industry that is accessible, understandable, and easy to perform for organisations with different capabilities, resulting in the improvement of evaluation programs carried out by those organisations, as well as improved outcomes for trainees.

By achieving this objective, the construction industry can improve the planning, implementation, and evaluation of upskilling programs, leading to more efficient use of resources, higher program effectiveness, and better outcomes for trainees. The development of a comprehensive evaluation framework will also contribute to the advancement of evaluation practices in the construction industry, which can benefit other industries and sectors as well.

4.3.3. Partial Design Requirements

The research findings and the pilots in the analysis phase provided essential insights, and informed which aspects should be considered primarily upon designing a solution. These first considerations serve to frame future choices in design, relating to the needs and context of evaluations.

a. **Clarification the evaluation rubric**. There is a need to clarify the structure of evaluation programs proposed by the rubric. Despite being based on the well-known Kirkpatrick's 4-level model, this relation does not seem clear to evaluators. The three levels proposed by the rubric and their different approaches, characteristics and objectives need to be clear

for designers of evaluations. Jointly, the subjective and objective dimensions of measurements should be explained and promoted.

b. **Comprehensible step-by-step guidance**. Evaluators must be supported in the designing of evaluation programs using the rubric. The guidance provided should be comprehensive and easy to understand, with clear instructions and easy-to-follow steps, including the selection of appropriate evaluation methods, the identification of evaluation objectives, and the development of evaluation plans. The guidance should also be designed to be applicable to all partnering organisations, regardless of their size or capabilities.
c. **Prevent unspecific and unmeasurable learning goals**. Organisations need guidance on how to define clear and measurable learning goals that align with their training objectives. This guidance should include clear instructions and easy-to-follow steps that can help organisations develop effective learning goals.
d. **Selection of survey tools in the rubric**. Organisations need to be assisted in selecting the most appropriate survey tools to measure the effectiveness of their training programs. This guidance should consider the different capabilities of partnering organisations in collecting and treating data obtained from survey.
e. **Avoid complex analysis of data**. The ability of organisations to perform higher-level statistical analysis of data is limited. The data obtained from evaluations should be descriptive and easily interpreted. The visualisation of data should be facilitated.

By fulfilling these partial design requirements, organisations will be able to design and implement evaluation programs that can accurately measure the effectiveness of their upskilling programs in the construction industry.

4.3.4. Initial Design Propositions

Design propositions refer to core ideas or hypotheses that serve as the foundation for the design process and are derived from the refinement of the understanding of the problem and context. Based on the information obtained in the analysis and exploration phase, it was

determined that extensive guidance had to be provided for organisations to be able to design evaluation programs. A roll of requirements and propositions was laid out to define the most suitable tools (see Figure 6).

Figure 6

Initial Design Requirements and Propositions

Partial design requirements	Initial design propositions
Clarification of the evaluation rubric.	Developing a framework for the evaluation rubric
Comprehensible step-by-step guidance.	Comprehensive and organised resource tool providing information, instructions, or guidance.
Prevent unspecific and unmeasurable learning goals.	Guide for structured formulation of learning goals
Selection of survey tools in the rubric.	Explanation and categorisation of survey instruments
Avoid complex analysis of data.	Prioritisation of the collection of data that can be easily visualised and descriptively analysed.

A. **Clarify the evaluation rubric.** The current rubric recommends methods and instruments that presuppose previous knowledge or familiarity with evaluation processes. This does not seem to be the case for evaluators in the construction industry. Hence, there is a need to scaffold and support inexperienced evaluators by providing them with the clarifications and knowledge required to design effective evaluations.

B. **Comprehensive and organised resource tool providing information, instructions, or guidance.** A design solution for this challenge must contemplate not only the clarification of the evaluation rubric, but also provide a structured way of carrying out the processes involved in the design and implementation of evaluations.

Due to the lack of experience of possible evaluators in the construction industry, there is a need for specific guidance that can be easily followed, in a stepwise approach.

C. **Guide for the structured formulation of learning goals.** The definition of learning goals is generally considered a systematic process (Chatterjee & Corral, 2017; Fessl et al., 2021; Marzano, 2009), containing defined structures Marzano (2009) suggests the use of SMART criteria for defining learning goals, englobing five components. This modular structure could be reproduced as a tool.

D. **Explanation and categorisation of survey instruments – Question Packs.** Organising survey questions plays a crucial role in enhancing the survey experience by simplifying readability and completion. It brings clarity to the concept area and context, as well as facilitating data analysis and ensuring optimal results when aligned with the evaluation objectives.

E. **Prioritisation of the collection of data that can be easily visualised and descriptively analysed.** Due to the limited capability of data analysis of evaluators in the construction industry, the collection of data should focus on methodologies that are easily transposed into visualisations and by using simple and accessible digital platforms.

5. Design and Construction

In the previous phase – Analysis and Exploration – the problem was defined, a long-range goal for the proposed solution was established and partial design requirements and propositions were specified. These results will inform and guide the actions to be carried out in the Design and Construction phase to create effective materials that support the use of the existing rubric in the design of evaluation programs. As posited by McKenney & Reeves (2018), this phase entails two processes. In *Design*, ideas on possible solutions are generated, considered, and checked in a systematic manner. In *Construction*, the prototypes are built and revised. In this study, prototypes were co-designed with stakeholders in two pilots and revised in two workshop sessions. The focus was on developing prototypes and testing them in real-world settings to gather feedback from participants. This feedback was then used to refine the developed designs and make any necessary changes to ensure that they effectively support organisations in achieving their evaluation goals.

To guide the exploration of various instantiations of design propositions and provide insights that could help organisations in the construction industry to develop effective evaluation, the following sub-research question was drafted:

> ***RQ.2*** *Which designs can support the autonomous development of evaluation programs for organisations in the construction sector?*

5.1. Methods of the Design Process

The following sections will outline the processes involved in the generation, consideration, and validation of ideas for the designs proposed in this study.

5.1.1 *Generating And Considering Ideas*

The initial design requirements and propositions established in the *Analysis and Exploration* phase of this study set the ground for generating ideas for possible design solutions. The initial ideas were further developed with the use of a morphological chart. A morphological chart is a structured tool used for exploring and generating multiple possible combinations or

variations of components within a system (McKenney & Reeves, 2018). The initial design propositions were reorganised and further developed, leading to broad, mid-level and specific propositions. The insights obtained during the analysis phase and the two piloting experiences led to the definition of two main propositions: the further development of the existing evaluation rubric into a framework, and the conception of appropriate tools to guide the design of evaluation programs (Figure 7).

Figure 7

Refinement of design propositions

Development of Evaluation Framework of Upskilling for the Construction Industry		
Broad propositions	**Mid-level propositions**	**Specific propositions**
1. Developing a framework for the evaluation rubric	Implement objective and subjective measurements	Self-reported assessment of learning goals
		Guidance on the design of knowledge and skill tests
	Clarify evaluation levels	Structuring different goals and objectives for each level
	Structure of evaluations	Determine different possibilities of outcomes for the design of evaluations
Instruction On the Design and Structuring of An Evaluation Program		
Broad propositions	**Mid-level propositions**	**Specific propositions**
2. Support the autonomous design of evaluation programs	Develop a referable instrument for ample usage	Create a Reference Guide
	Facilitate the definition of learning goals	Create a tool for defining learning goals
	Assist the selection of appropriate survey instruments	Classify, organise, and clarify survey instruments according to their objectives. Create ready-to-use replicable survey instruments
	Support data collection and analysis	Select appropriate technological tools

These two propositions yielded three design solution (DS) ideas:

5.1.1.1 Generating Design Solution 1: Creation of a Reference Guide.

The existing documentation in the project and the proposed evaluation rubric prescribe methods and instruments for the design of evaluations, but do not offer instructions on how to operationalise them. To cope with the challenge of providing a structured and systematic approach to the design of evaluations and the use of the rubric, these prescriptions need to be facilitated by instructional tools. Manuals and reference guides provide crucial information, instructions, and guidance. They offer clear explanations, step-by-step instructions, and troubleshooting tips, making them valuable references for users. By condensing extensive knowledge into user-friendly formats, manuals enable individuals to navigate complex tasks with confidence, efficiency, and optimal performance. Given the limited expertise of evaluators in the construction industry in designing structured evaluations, a manual would represent a permanent source of reference, in which processes are clearly described and explained. This idea, despite belonging to the proposition of supporting the autonomous design of evaluation programs, contemplates also the first proposition of developing a framework for the evaluation rubric.

5.1.1.2 Generating Design Solution 2: Creation of a Tool for Formulation of Learning Goals.

The formulation of clear learning goals was a consistent challenge during the piloting observations in the previous phase of the study. The SMART criteria (Doran, 1981) assure that learning goals are clear and measurable, whereas the use of a taxonomy such as Bloom's (Chatterjee & Corral, 2017) provide with the adequate verbs for their construction. However, during piloting, the operationalisation of both (SMART criteria and taxonomy) presented as a challenge for evaluators, as the added cognitive load of using two instruments and lack of experience often resulted in incomplete or unmeasurable statements. To make up for this shortcoming, a possible solution could be the development of a specific tool for the formulation of learning goals, in which evaluators would be required to comply with all the elements required for the formulation of learning goals. This could be accomplished in digital format, which would facilitate the task and limit errors. Spreadsheets,

for example, allow users to fill in data in an organised and structured manner, and are used for several purposes.

5.1.1.3 Generating Design Solution 3: Structuring of survey items.

The rubric provided evaluators with valid instruments for surveying different aspects related to first three levels of evaluation analogous to the Kirkpatrick 4-level model (Kirkpatrick & Kirkpatrick, 2006). However, valid instruments often comprise multiple related items which support the consistency of the construct (e.g., satisfaction with training, quality of teaching, normative success, etc.). However, the Analysis and Exploration phase revealed that evaluators in the construction industry lack knowledge and experience in the design of evaluations and surveys. Thus, evaluators need to be instructed on the utility and operationalisation of survey instruments. Furthermore, they need to be assisted in the collection and treatment of the data from surveys (e.g., measuring Likert scale items and compiling the data). Technological affordances in inquiry practices can facilitate several of these processes, such as the organisation and implementation of surveys, collection, and treatment of data.

5.1.2 Checking Ideas

One of the methods suggested by McKenney & Reeves (2018) for checking ideas in DBR is by using logic models. By allowing the visualisation of ideas, logic models help articulate the fundamental assumptions and operational processes associated with them. One of the simplest models is structured in five components (Frechtling, 2015): ***inputs*** (the resources invested in an activity), ***activities*** (the specific components to achieve desired outcomes), ***outputs*** (the direct and tangible products of the activities), ***outcomes*** (the desired changes or effects that occur as a result of the activity), and ***impacts*** (the long-term and broader effects or benefits of the project). The logic model for each idea is detailed in Appendix D.

5.1.2.1 Checking Design Solution 1: Creation of a Reference Guide.

The logic model outlines the development of a comprehensive and organised resource tool in the form of a manual to provide information, instructions, and guidance for designing structured evaluations in the construction industry. The *inputs* include existing project documentation,

evaluation rubric, and expertise in evaluation design and construction industry knowledge. The *activities* involve gathering and condensing relevant information, organising it into a user-friendly format, and developing the manual with expert input. The *outputs* are the manual itself, serving as a valuable resource tool for evaluators. The anticipated *outcomes* include increased accessibility to knowledge, improved understanding and utilisation of the evaluation rubric, enhanced ability to design structured evaluations, and increased confidence and performance of evaluators. The *impacts* are expected to be improved quality and consistency of evaluations, better assessment of construction projects, increased use of evidence-based decision-making, and overall effectiveness and success of construction projects.

Thus, the development of a manual holds great promise for improving evaluation design in the construction industry. By providing accessible information and stepwise instructions, the proposed tool can enhance evaluators' skills, confidence, and performance. The manual represents a valuable resource for evaluators, enabling them to navigate complex tasks with efficiency and optimal performance. By bridging the gap between existing project documentation and the evaluation rubric, the manual facilitates a structured and systematic *approach to evaluation design in the construction industry.*

5.1.2.2 Checking Design Solution DS 2: Creation of a Tool for Structured Formulation of Learning Goals

. The creation of a digital tool or guide to help with the structured formulation of learning goals addresses challenges observed in the previous study phase. The *inputs* include knowledge and expertise on goal formulation, the SMART criteria, Bloom's taxonomy, and feedback from evaluators. The *activities* involve researching effective strategies, developing a guide/tool integrating the SMART criteria and Bloom's taxonomy, designing it in a digital format, and refining it through testing. The expected *outcomes* include improved understanding and application of goal formulation principles, reduced errors and cognitive load, and increased confidence for evaluators. The *impacts* of implementing this tool are anticipated to be improved learning goal quality, better alignment between goals and

instructional activities, enhanced effectiveness of learning interventions, and improved student learning outcomes.

The development of a digital tool or guide for the structured formulation of learning goals presents a promising solution to the challenges identified in the previous study phase. By incorporating the SMART criteria and Bloom's taxonomy, the tool provides clear guidance and facilitates the creation of comprehensive and measurable learning goals. Its digital format offers convenience and reduces errors, therefore simplifying the process for evaluators. By improving the formulation of learning goals, the tool has the potential to enhance its quality and effectiveness, leading to better alignment between instructional activities and assessments.

5.1.2.3 Checking Design Solution 3: Structuring of Survey Items. The use clarification, organisation, and facilitation of survey instruments in evaluations can potentially enhance their use and adoption by partners in the project. The inputs include valid survey instruments, feedback from the Analysis and Exploration phase, and technological affordances. The activities involve categorising the survey instruments, developing explanations and instructional materials, creating guidelines for data prioritisation, and incorporating technology for survey organisation and data treatment. The outputs consist of categorised survey instruments, instructional materials, prioritisation guidelines, and technological tools. The anticipated outcomes include increased knowledge and understanding among evaluators, improved instrument selection and prioritisation, enhanced data collection and treatment skills, and efficient use of technology. The impacts are expected to be enhanced quality of evaluation data, improved visualisation and analysis of survey data, evidence-based decision-making, and increased evaluator confidence and competence.

The design solution three offers a comprehensive approach to enhancing the use of survey instruments in evaluations. By categorising the instruments, providing explanations and instructional materials, and establishing prioritisation guidelines, evaluators can gain a

better understanding of the purpose and application of different surveys. This leads to improved selection and prioritisation of instruments based on evaluation needs and desired outcomes. The incorporation of technology accelerates survey processes, enabling evaluators to visualise and analyse data effectively, make evidence-based decisions, and build confidence in their evaluation skills.

5.2 Prototyping

Once ideas have been selected, considered, and assessed for their feasibility, considering all design requirements, the prototypes could be built. The process of prototyping involved creating a preliminary version and checking its functionality, design, and user experience. Once the prototypes were created, it was tested and evaluated to gather feedback, identify potential improvements, and make necessary iterations. This iterative process continued until a satisfactory prototype was developed, serving as a basis for further development or implementation.

5.2.1. Prototype Design Solution 1: Reference Guide (RG).

The RG should comply with the initial proposition of being a comprehensive and organised resource tool that provides information and instructions (relating to design requirements: *Clarification the evaluation* rubric and *Comprehensible step-by-step guidance*). The first step into the creation of such tool is to map and organise the information it should contain. To increase potential effectiveness, this study adopted techniques supported by literature. The minimalist approach was adopted for the design of the RG. Van der Meij and Carroll (2003) set four principles for designing minimalist instruction: choosing action-oriented approaches, anchor the tools in the task domain, support error recognition and recovery, and support reading to do, study and locate. Farkas (1999) highlights the importance of using a stepwise approach, investing on visual aids and suiting the rhetoric to the target audience.

Evaluation123. The process of developing the evaluation rubric into a structured framework started with adopting a specific name for it. The purpose of naming the framework was to make it easier for potential users to refer to it and facilitate communication. The term Evaluation123 seemed appropriate, as it not only bears a mnemonic to the three levels of evaluation (see Figure 8) adapted from the Kirkpatrick's model, but also conveys and idea of stepwise approach to the design of evaluations, which had been mentioned as a design requirement during the pilots in the analysis and exploration phase. It is reasonable to assume that this could be positively perceived as a simpler or easier process, even for inexperienced evaluators.

Figure 8

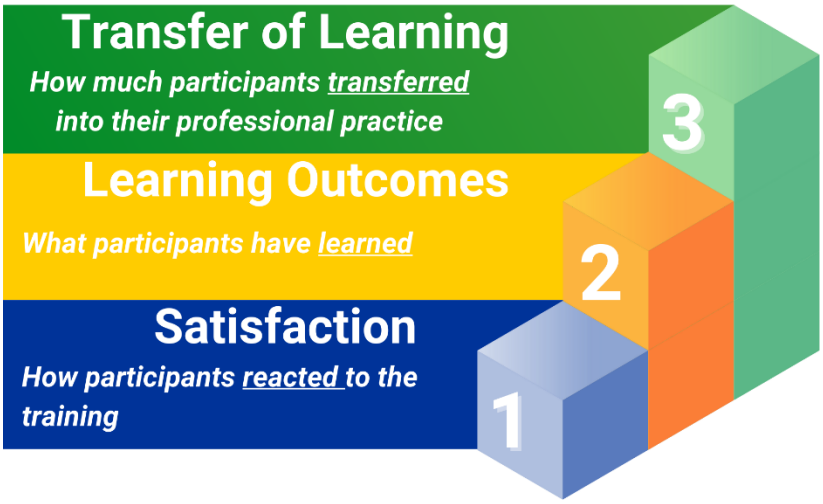
Logo for Evaluation123

Clarification of the different evaluation levels. As it was revealed in the analysis phase of the study, organisations do not see evaluations as part of the core processes in their training efforts. As a result, evaluations tend to be experimentally designed and implemented, which reflects on the quality of the results of evaluations. Ultimately, these results may not be sufficiently informative or descriptive of the effectiveness the trainings. Attaining evaluation goals is only possible when those are clearly set and described. Therefore, the goals and objectives of each level in the rubric needs to be clarified and explained. A chapter in the RG

was added to address and explain the different levels (*Reaction, Learning Outcomes* and *Transfer of* Learning) and their respective objectives. The explanation was accompanied by proper imagery and visual clues to assist comprehension of the levels and how they are structured in a ranked manner (see Figure 9).

Figure 9

Evaluation123 3-Level Model



Implementation of objective and subjective measurements. Objective measures in assessments are criteria that can be quantified or measured, and are utilised to assess performance, knowledge, skills, or desired outcomes in a standardised and impartial way. Conversely, subjective measures in assessments refer to qualitative or judgment-based criteria used to evaluate performance, skills, or other desired outcomes based on the subjective interpretation or opinion of an individual. These measures are typically based on personal judgment, impressions, or perceptions. Unlike objective measures, subjective measures rely on the individual's subjective interpretation rather than quantifiable data.

The rubric for evaluations of BUSLeague proposed that both dimensions (objective and subjective measurements) should be used in assessments, as means of increasing the

accuracy of evaluations. This approach is supported in literature and both measurements are considered complementary to each other (Merchant et al., 2010).

Kirkpatrick's model traditionally implies measurements in only one dimension. For level 1 (Reaction), self-reported measurements of satisfaction are surveyed, and for the following levels, the model adopts objective measurements. The rubric prescribes that the measurements of both dimensions should be enabled at each level (see Figure 10).

Figure 10

Comparison of Kirkpatrick's model and the Evaluation123

		Kirkpatrick's 4-level model	**Evaluation123**
Level 1	**objective**		**Design Principles**
	subjective	Reaction	Satisfaction
Level 2	**objective**	Knowledge / Skill Test	Knowledge / Skill Test
	subjective		**Learning Goals**
Level 3	**objective**	Knowledge / Skill Observation	Knowledge / Skill Observation
	subjective		**Learning Goals / Motivation**

The rubric selected instruments that could be used as valid instruments for assessing upskilling activities. For level 1 of the rubric (Satisfaction), it is suggested the use of Saks and Belcourt's (2006) instrument for assessing design principles as an objective measurement. At level 2 (Learning Outcomes), the rubric suggests using Bloom's taxonomy (Oliver et al. 2004) as a guideline for knowledge and skills assessments. The taxonomy is at the core of the definition of learning goals. Thus, based on evidence from literature (Andrade, 2019; Arico et al., 2018; Pisklakov, 2014), this study suggests the use of learning goals for self-reported assessments, adding a subjective dimension for assessments on level 2. For

level 3 (transfer of learning), the rubric suggests the use of Chauhan et al. (2017) instrument for measuring motivation and self-perceptions on trainings. Studies have shown that self-assessments potentially impact performance, self-efficacy, and motivation of learners (Rolheiser & Ross, 2001). Sedikides and Strube (1997) found that self-knowledge certainty can be enhanced, while Pisklakov (2014) emphasised it as a learning opportunity. Thus, this study also suggests the assessment of learning goals as self-reported measurements when assessing transfer of learning, alongside the prescribed post-observation of knowledge and skills in Kirkpatrick's model.

To combine objective and subjective measurements by the recommendations of the European Framework for Qualifications[7], the Evaluation123 must utilise a variety of assessment methods. It is important to ensure that the learning goals of the self-reported assessment and the goals of the knowledge and skill tests are aligned. For example, if learners report that they have achieved a certain learning goal, the knowledge and skill tests should reflect this level of achievement. Similarly, if the knowledge and skill tests indicate that learners have not yet achieved a certain level of proficiency, this should be reflected in their self-reported assessment of learning goals.

Designing knowledge and skill tests is an objective assessment method that involves creating evaluations that measure specific knowledge or skills. These tests can take many forms, including multiple-choice, short-answer, or performance-based assessments. By using standardised and well-designed tests, it is possible to ensure that all learners are evaluated based on the same criteria, regardless of their personal biases or opinions.

Overall, combining objective and subjective measurements can provide a more comprehensive and accurate assessment of learners' knowledge and skills. By utilising methods such as self-reported assessment of learning goals and design of knowledge and

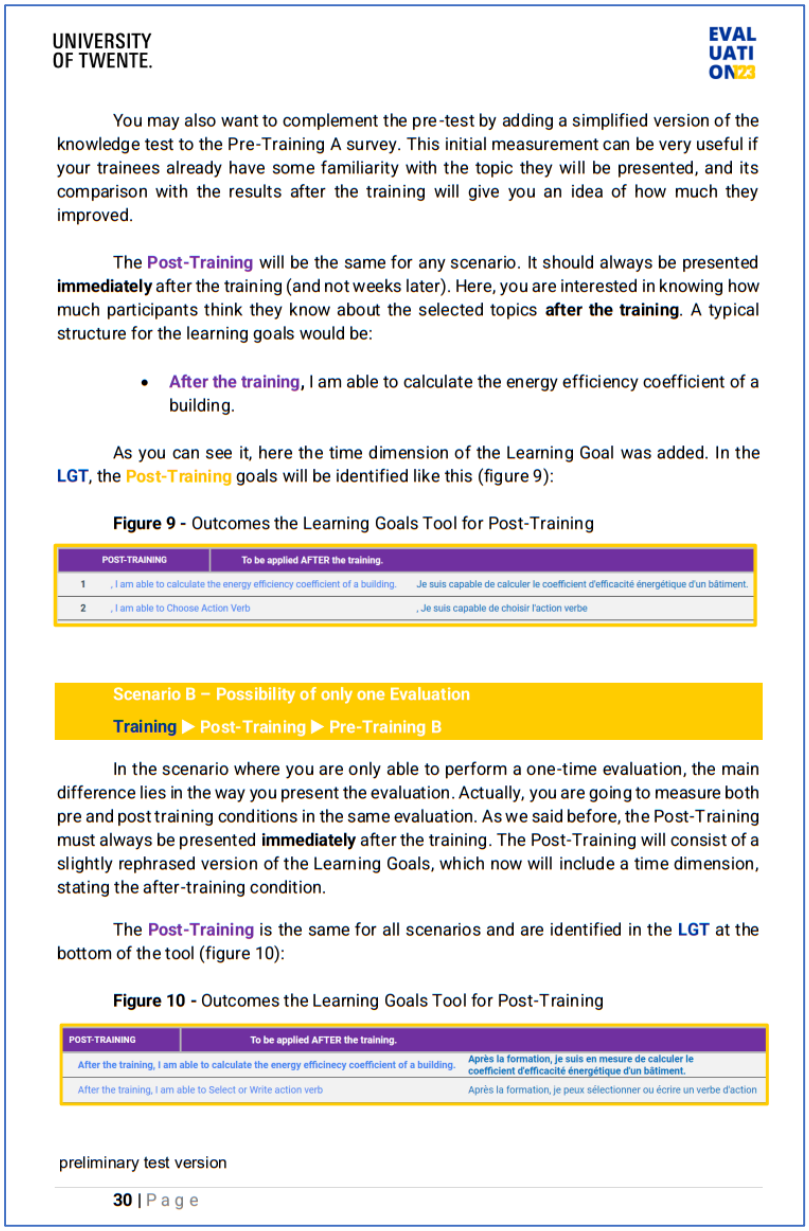
[7] https://europa.eu/europass/en/europass-tools/european-qualifications-framework

skill tests, the Evaluation123 can ensure that learners are evaluated based on both their perceptions and standardised criteria.

Structure of evaluations. According to Larson and Berliner (1983), evaluations can be conceptualised as a system consisting of three primary components: inputs, processes, and outcomes. Thus, it is important for evaluators to understand these three components, their interrelationships and how to operationalise them (see example in Figure 11).

Figure 11

Sample Page of the RG



UNIVERSITY OF TWENTE.

EVAL UATI ON23

You may also want to complement the pre-test by adding a simplified version of the knowledge test to the Pre-Training A survey. This initial measurement can be very useful if your trainees already have some familiarity with the topic they will be presented, and its comparison with the results after the training will give you an idea of how much they improved.

The Post-Training will be the same for any scenario. It should always be presented **immediately** after the training (and not weeks later). Here, you are interested in knowing how much participants think they know about the selected topics **after the training**. A typical structure for the learning goals would be:

- **After the training,** I am able to calculate the energy efficiency coefficient of a building.

As you can see it, here the time dimension of the Learning Goal was added. In the LGT, the Post-Training goals will be identified like this (figure 9):

Figure 9 - Outcomes the Learning Goals Tool for Post-Training

	POST-TRAINING	To be applied AFTER the training.
1	, I am able to calculate the energy efficiency coefficient of a building.	Je suis capable de calculer le coefficient d'efficacité énergétique d'un bâtiment.
2	, I am able to Choose Action Verb	, Je suis capable de choisir l'action verbe

Scenario B – Possibility of only one Evaluation

Training ▶ Post-Training ▶ Pre-Training B

In the scenario where you are only able to perform a one-time evaluation, the main difference lies in the way you present the evaluation. Actually, you are going to measure both pre and post training conditions in the same evaluation. As we said before, the Post-Training must always be presented **immediately** after the training. The Post-Training will consist of a slightly rephrased version of the Learning Goals, which now will include a time dimension, stating the after-training condition.

The Post-Training is the same for all scenarios and are identified in the LGT at the bottom of the tool (figure 10):

Figure 10 - Outcomes the Learning Goals Tool for Post-Training

POST-TRAINING	To be applied AFTER the training.
After the training, I am able to calculate the energy efficinecy coefficient of a building.	Après la formation, je suis en mesure de calculer le coefficient d'efficacité énergétique d'un bâtiment.
After the training, I am able to Select or Write action verb	Après la formation, je peux sélectionner ou écrire un verbe d'action

preliminary test version

30 | Page

The design of an evaluation program will lead to different results depending on the inputs (resources and methodologies) provided at the beginning of the process. The evaluation for an upskilling activity that aims at a certification may be considerably more complex than one destined to simply inform trainers on the learning outcomes. Likewise, the availability of time, material and human resources may impose decisions that will shape the evaluation program. Therefore, evaluators need to be aware of the factors that might influence their efforts and use this information to manage them satisfactorily.

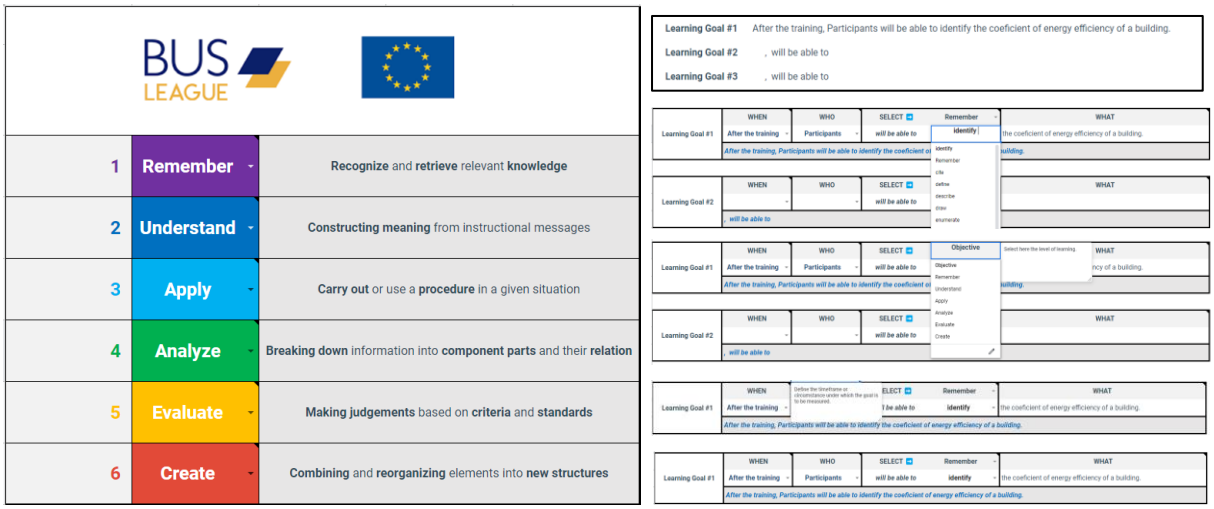
The prototype version of the RG was a 19-page document containing instructions on how to structure evaluations (pre-tests, post-tests) and an explanation of the three levels and their assessment tools. The RG already embedded an explanation of the use of the LGT and referred to the prototype of the QP and how to use them in digital platforms.

5.2.2. Prototype Design Solution 2: Learning Goals Tool (LGT)

In the first iteration of pilots, the definition of learning goals was deemed as one of the most complex and time-consuming activities in the evaluation design process (related to design requirements *Prevent unspecific and unmeasurable Learning Goals*). Therefore, this study proposed the development of a specific instrument for designing effective learning goals. The design and evaluation of learning goals is an essential process in ensuring effective learning outcomes (Fessl et al., 2021; Marzano, 2009; Scriven, 1994). To address this challenge, this study proposed the development of a specific instrument for designing effective learning goals, the “Learning Goal Tool” (LGT; Figure 12).

Figure 12

Screenshots of the Learning Goals tool v.01[8]



The LGT instrument helps to design learning goals systematically. It is built on an online spreadsheet (Google Sheets), in a form-like structure that is accessible and familiar to most users.

The structure for designing learning goals was based on the SMART criteria (Doran, 1981) . These criteria establish that learning goals should be specific, measurable, attainable, relevant and time-bound. Thus, the syntax for learning goals statements included the following elements (see Figure 13).: *When (time-bound), Who (specific), Action Verb (measurable)* and *What (attainable).* The relevance criterion underlines the ensemble of all elements and characterises a sell-defined learning goal.

Figure 13

Syntax of Learning Goals in the LGT

	WHEN	WHO	SELECT	Remember	WHAT
Learning Goal #1	By the end of the session	Trainees	*will be able to*	describe	the energy labels of a building
	By the end of the session, Trainees will be able to describe the energy labels of a building				
Spanish	*Al final de la sesión, los alumnos podrán describir las etiquetas de energía de un edificio*				

[8] https://docs.google.com/spreadsheets/d/1WK313c12spaPAgE_g6jxE55igm1s6D0_HLblN2eArPI/edit?usp=sharing

Action verbs are selected from Bloom's taxonomy for levels of learning, namely remember, understand, apply, analyse, evaluate, and create. Each level has a descriptor and a drop-down box with related concepts and cognates. The levels are numbered and coloured, providing a clear hierarchy of learning.

The instrument includes ten learning goal (LG) rows, each with several fields for defining the learning goal. The first field (WHEN) defines the time frame or circumstance under which the goal is to be measured, with sample phrases such as after the training/course, by the end of the workshop, or at the end of the session. The second field (WHO) identifies the target group of the training, with sample subjects including trainees, students, learners, participants, and consultants.

The third field is a select arrow pointing to the objective field, which includes the fixed locution "will be able to" as an enabler. The objective field provides a drop-down box with six levels of learning(Adams, 2015b), and a second drop-down box with a list of verbs attaining the specified level. The fourth field (WHAT) is an open field for describing the knowledge, skill, or competence to be achieved.

The development of a learning goal design instrument offers a practical solution for designing effective learning goals. The instrument's form-like structure, Bloom's revised taxonomy, and clear hierarchy of learning provide a user-friendly interface for defining learning goals.

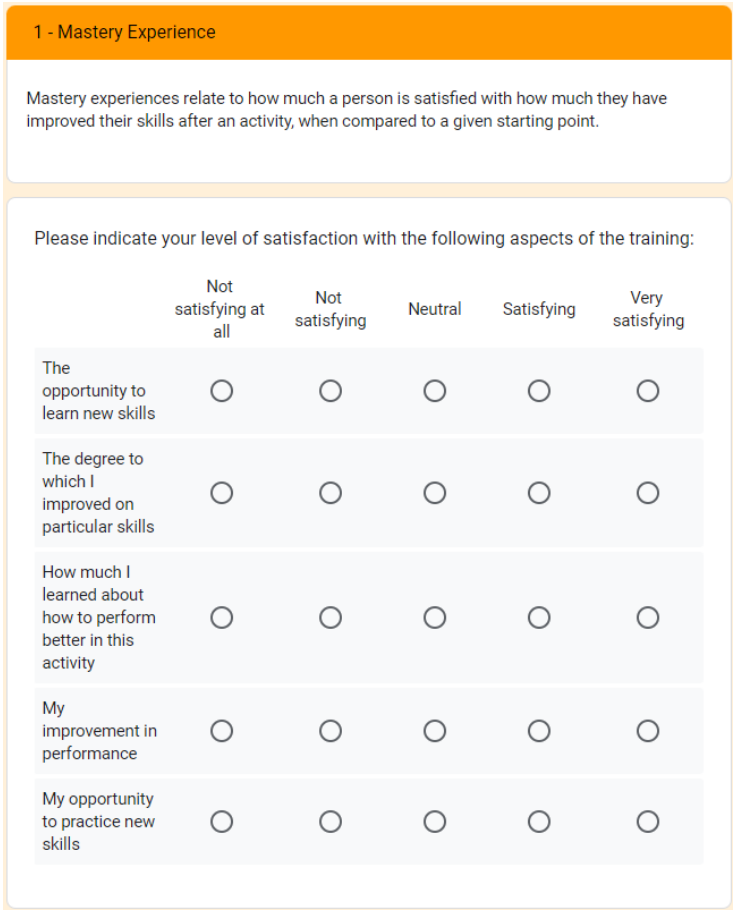
5.2.3. Prototype Design Solution 3: Question Packs (QP)

During the analysis phase, it was detected that the participating organisations were not using the rubrics with validated survey instruments for evaluation because their use, intention and significance were not presented. Validated instruments are generally more elaborated and extensive than the experimental survey tools developed internally by organisations, because they require a more complex set of items to assure the reliability and validity of measurements.

To encourage its adoption and facilitate their use (related to design requirement: *Selection of survey tools in the rubrics)*, the rubrics were reorganised in question packs, and presented as modular packs that could be combined (see example in Figure 14).

Figure 14

Example of survey instrument organised as digital Question Pack



With this electronic tool, users can create their surveys and evaluations by importing the packs they deem more relevant to their interests and purposes. A succinct and easy explanation of the constructs underlying each instrument was added to each pack and further developed in the reference guide.

This study used Google Forms because it is a more accessible platform and is compliant with the european General Data Protection Regulation GDPR[9]. Google Forms also exports results automatically in spreadsheet format, which allows the user to produce simple visualisations of data and run basic descriptive functions. A few of these possibilities were demonstrated as examples during the explanatory section of the activity.

5.3. Construction - Piloting of Reference Guide, Learning Goals Tool and Question Packs

In the construction process of the Design and construction phase, the prototypes were trialled with partners in two pilot sessions with individual partners, and two workshops with a larger audience in the project.

5.3.1. Participants in Piloting of Design of Evaluations

Two pilot interventions were implemented (see Figure 15). In the third pilot, an evaluator from an Austrian energy agency carried out an evaluation for a micro training presented in a videoconference. The second pilot involved an evaluator from a public-interest organisation assessing skills of practising professionals from Spain. These evaluators were professionals entrusted with overseeing the training initiatives within their respective organisations.

Figure 15

Overview of Activities Carried Out in the Design and Construction Phase

[9] https://eur-lex.europa.eu/eli/reg/2016/679/oj

5.3.2. Methods of Piloting Design of Evaluations

A similar procedure was used regarding the pilots in the analysis phase. However, this time the prototypes of tools were used to assist the design of evaluations. Semi-structured interviews focused on the process of building and implementing an evaluation program using the proposed tools, as well as identifying which processes are perceived as more challenging for organisations. Debriefing sessions also focused on the usefulness and usability of the proposed tools, and participants were requested to provide feedback regarding their execution of an evaluation design and their perception of the tools provided for this purpose. Please refer to Appendix C (items 27-52) for the specific items.

The answers provided to the semi-structured were later coded using the same scheme developed for first pilots. This enabled the exploration of areas of knowledge that were not addressed in previous iterations, enhancing comprehension of organisational needs, limitations, and capabilities when designing evaluation programs.

5.3.3. Integrated Results of the Pilots 3 and 4 in the Design Phase

During pilots 3 and 4, a prototype of the RG, LGT and QP were tested during the design of evaluation programs. In pilot three, the tools were used in a reiteration of the first pilot, whereas in pilot 4, the tool was used to define learning goals for an evaluation that aimed at the recognition and certification of competencies. The interviews carried out in the debriefing sessions aimed at understanding the shortcomings and opportunities concerning the designed tools, and to determine which improvements needed to be implemented.

5.1.3.0 Results from Piloting the Reference Guide. During the pilot phases, the reference guide was acknowledged as a valuable resource for the participants. While the guide contained valuable information, its complexity made it difficult for the participants to fully comprehend and extract actionable insights for enhancing the evaluation process. The integration of explanations of the LGT and QP was considered useful, consolidating the idea of an integrated and comprehensive tool for assisting the design of evaluations. However, more aspects of evaluations, such as different scenarios (i.e., pre and post-tests) and the

different dimensions of evaluations (i.e., objective, and subjective measurements) require further explanation and argumentation.

5.1.3.1 Results from Piloting the Learning Goals Tool.

Overall, the LGT tool was appreciated and considered satisfactory by the respondents. The LGT was reported to set a clearer structure for the definition of LGs, assuring the presence of the main elements and narrowing their scope to single, measurable actions. Secondarily, the use of SMART LGs throughout current and previous pilots has enabled respondents to perceive the potential for increased comparability across interventions.

However, the use of Bloom`s taxonomy (Bloom, 1956/1984) was deemed too complex by the respondents. Determining the level of cognitive processing requires an understanding of the conceptualisations. This potentially limits the use of the tool, if the user is not able to comprehend what cognitive processes are and their different stages. The list of verbs offered alongside each level was also considered too extensive. While navigating through the list, respondents affirmed the list to be overwhelmingly long, and sometimes unsure of the clear definition of a given word. It was also mentioned that trainings in the construction sector are largely focused on building practical skills. These skills are not completely contemplated by Bloom's taxonomy, which focuses on cognitive processes.

Therefore, the use of the LGT needs to be facilitated by reducing the steepness of the learning curve required to determine the level of learning and the selection of adequate verbs. It should also keep a clear and concise language, to accommodate users with varying levels of proficiency in the English language.

5.3.3.1. Results from Piloting the Question Packs.

The use of the QP was mentioned to be "easy" and "intuitive." Its organisation in modulable packs was well perceived and their use did not present major challenges, as participants were already familiar with the adopted platform (Google Forms). After the implementation of the training, evaluators were able to easily collect the data, as the platform automatically compiles all

data in a spreadsheet format. This allowed them to produce reports describing and analysing the results.

5.3.4. Design improvements resulting from Pilots 3 and 4

In the next subsections, the improvements implemented in the tools as a result of the piloting sessions will be presented.

5.3.4.1. Improvements in the Reference Guide. Based on the feedback from the piloting, several improvements have been made to the RG. The language in the RG was simplified and adapted to suit a broader audience, in line with the recommendations from Farkas (1999). To increase comprehension, examples were added whenever explaining a procedure or concept. The RG also incorporated extensive imagery to support understanding of the concepts, and also meant as visual aids for assisting the use of the online tools (LGT and QP). Lastly, the structure of the chapters was modified, and the three distinct levels were made more noticeable, differentiating colours and adding a clearer separation of objective and subjective measurement processes.

5.3.4.2. Improvements in the Learning Goals Tool. To simplify the process of determining learning goals, this study sought to understand the type of learning goals that were being developed during the pilots. In total, for all four pilots, twenty-six different learning goals were drafted (see Figure 16).

Figure 16

LGs created during pilot sessions.

Level of learning	Pilot 1	Pilot 2	Pilot 3	Pilot 4	Total of LGs
Remember	3	5	2	1	11
Understand	-	5	-	-	5
Apply	1	-	1	6	8
Analyse	-	-	-	-	-
Evaluate	1		1		2
Create	-	-	-	-	-

The pilots showed an indication of a tendency for the definition of learning goals at the least complex levels. A possible explanation for this is that short upskilling interventions in the construction sector, such as micro trainings, may not aim at higher levels. Another possibility is that the contents of upskilling trainings in energy efficiency for the sector may focus more on the acquisition of knowledge on a topic, and at a certain level on its practical application. In pilot four, the higher concentration of LGs in the apply level is due to the objective of the evaluation, which was to determine the proficiency in skills of practising workers rather than the gained knowledge or skill after a training.

Based on the knowledge gained in the pilots, three main meta-categories can be identified within Bloom's taxonomy:

- **Acquisition of knowledge**: referring to the acquisition, internalisation, and retrieval of conceptual knowledge.
- **Application of knowledge**: implying the ability to apply or use the acquired knowledge in a context.
- **Critical use of knowledge**: involving the ability to critically analyse or evaluate processes, and to apply knowledge and skills in different or innovative ways.

Therefore, this study proposed to simplify Bloom's taxonomy into three categories, by grouping cognitive levels that relate to the same meta-category (Figure 17).

Figure 17

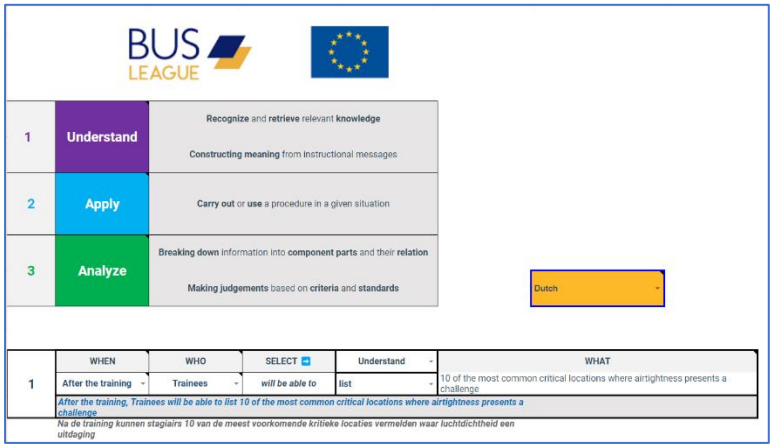
Simplification of Bloom's Taxonomy

	Meta category	**Bloom's cognitive levels**
Understand	Acquisition of knowledge	remember, understand
Apply	Application of knowledge	apply
Analyse	Critical use of knowledge	analyse, evaluate, create

In the LGT, an explanation of the overarching themes about each level was added to increase comprehension and assist in choosing other category and verbs. The list of verbs was also redesigned to reflect the changes in the categories. The number of verbs in each list was also significantly reduced by eliminating synonyms. In addition, the verb list contained in the tool was redesigned to encompass the verbs that reflect each condensed level. This reduction was achieved by eliminating synonyms and uncommonly used verbs (Figure 18).

Figure 18

Learning Goals Tool v.02



5.3.4.3. Improvements in the Question Packs. The QP required minor changes, mostly involving the clarification and facilitation of use. Each module, which correspond to the measurement of a single construct, received a proper definition as a subheading, visible to evaluators. Naming of modules was also adapted to disambiguate homonymous items (e.g., Satisfaction in Cunningham, 2007; So & Brush, 2008; and Athiyaman, 1997).

5.4. Workshop 1: Trialling the Reference Guide, Learning Goals Tool and Question Packs

Workshops were also carried out twice during the design phase, where participants were introduced to the tools and asked to perform evaluation-design activities in dyads or smaller groups to test the usability of the tools. At the end of each session, a guided, semi-structured interview was conducted to gather relevant opinions, perceptions and feedback on the activity and tools. The interview followed a similar form to the one used in debriefing sessions for pilots in phase two, in which participants were asked to report on their experience in executing an evaluation design and their perception of the tools offered for this purpose (see Appendix C, items 27-52). The aim was to determine which features were the most helpful and which should be further developed to make them accessible and facilitate the process of designing an evaluation.

The interviews were recorded with consent and transcribed using Amberscript, with any interactions in other languages translated into English. The software ATLAS was used to code the interviews. These sessions aimed to determine the most helpful features of the tools and to identify areas for further development to facilitate the process of designing an evaluation.

5.4.1 Participants of Workshop 1

A total of nine participants from eight different organisations took part in the workshop. All participants were appointed as responsible for training in upskilling activities in their organisations during the BUSLeague project. Before the activity, participants received a mock case study, describing a fictitious organisation and their training program and their need for evaluation.

5.4.2 Method for Workshop 1

The purpose of the first workshop was to present the preliminary version of the designed tools and assess their feasibility to the consortium of partners in the BUSLeague

project. The preparation material included questions meant to prompt reflection on the process of designing evaluations and their challenges. The overall aim of the workshop was to support the autonomous application of the evaluation framework and test the LGT and use of question packs based on the rubrics.

The workshop took place via a video conference. The sessions were recorded and later transcribed using Amberscript. The activity performed was structured in two segments and participants were grouped into four teams. At the end of the proposed time for each activity, participants were convened together and interviewed as focus groups (see Appendix E). The structure of the interview allowed the identification of the tools' strengths, weaknesses, opportunities, and threats (SWOT)

The first available version of the RG was presented during the workshop. As a comprehensive reference guide for evaluations, its use and utility could not be assessed during the event, but its features were demonstrated and explained. Thus, the focus of the first workshop was trialling the LGT and the QP.

The first activity focused on the use of the LGT. Participants were required to define appropriate learning goals, using the prompts given in the preparation material and the LGT. During the activity, participants were observed and clarification on the use of the tool was provided upon request. The second segment of the workshop was devoted to the use of the question packs to develop a reaction survey (level 1) and the self-reported assessment of learning goals (level 2). Participants were requested to use the modules to assemble a survey and assessment adequate to the case study used in the activity. Feedback was collected to inform improvements in design.

5.4.3. Results of Trialling the Learning Goals Tool and Question Packs in Workshop 1

5.4.3.1. Results from Trialling the Learning Goals Tool. All teams were able to make an intuitive use of the tools, with minimal support. This may be attributed to the choice of the spreadsheet format, which was familiar to all users. At the end of the activity, all

teams were able to elaborate on at least two learning goals out of four possibilities. This was in line with the limited time offered for the activity. It was also mentioned that the tool is accessible and simple, with a positive reaction from all teams. Conversely, the selection of adequate verbs seemed to still present as a challenge to users. The list of verbs was drafted from Bloom's taxonomy, which is highly focused on cognitive skills. When trying to define learning goals that involved practical skills, participants had difficulties finding adequate verbs. The reasons underlying this difficulty may be the fact that users might be trying to find in the lists an exact match for the verbs they intended, instead of writing down their options. The list of verbs should be seen as a reference list, and not as a limited array of possibilities.

When asked about the usefulness of the LGT, participants declared it to be useful. It was pointed out that the tool clarifies what the main elements of a learning goal are in a structured manner ("*I think it's very nice way to make it accessible for people to set up learning goals")*. The added explanation of the learning levels was also deemed useful, as it frames the definitions and assists the selection of action verbs ("*I think one of the most important thing is that you classify the groups from understand, apply and analyze to the specific words"*).

Participants reported that the LGT helped them to have a clearer idea of their evaluation goals. They understood the purpose and objective of evaluation, rather than seeing it as an externally compulsory requirement. Evaluation and learning goals are interdependent, as evaluation aligns with learning goals, provides feedback, guides both the instruction and the achievement the desired outcomes. The tool also enabled them to peek into Level 2 evaluation, which they usually did not do. However, some participants expressed uncertainty about the difference between Level 1 and Level 2 and what their evaluation goals might be. The participants suggested an increase in the focus on evaluating trainers.

Regarding the usability of the tool, some participants found it challenging to understand the levels in the taxonomy. Nevertheless, other participants found that the organisation of the tool helped them to clarify their evaluation goals and understand the

application of the rubric. Participants suggested that simplification of items could make the tool easier to use. It was observed that the automatization of certain tasks (e.g., filling of common fields to all learning goals) could facilitate the use of the tool.

The analysis of the feedback provided by the focus groups identified the following SWOT for the LGT. The strengths of the LGT were identified as assisting with clarifying evaluation goals, understanding the application of the rubric, and helping assess L2. The weakness of the LGT was identified as being unclear on the differences between the levels of learning as described in the taxonomy, and users seemed unsure of what to do during the activity. The opportunity identified for the LGT was to simplify the items to make it easier for them. The threat identified was that not all users have a clear view of evaluation goals, which limits the use of the tool, and the difficulty in designing pre and post-tests.

In conclusion, the LGT was effective in assisting users in clarifying Learning goals. However, the simplification of the LGT items could make it easier to use and understand. The tool's use is limited by the fact that not all users have a clear view of evaluation goals.

5.4.3.2. Results from Trialling the Question Packs. The use of question packs did not raise many difficulties. However, participants still questioned the need for multiple items when measuring one construct (i.e., items increasing validity and reliability). The length of the final surveys seemed to limit the evaluators' willingness to adopt more extensive surveys that measure more constructs. It was also commented that evaluators would normally add more items related to the assessment of teaching, rather than measuring constructs related to self-efficacy or normative success. These perceptions might have been influenced by the fact that participants did not have time to read the RG during the workshop, and were unfamiliar with the explanations it contained. The QP also proved itself useful in facilitating

the design of assessment surveys, and the modular aspect rendered it more accessible to a broader audience of evaluators.

5.4.4. *Design improvements resulting from Workshop 1*

The following subsections explain the improvements made to the proposed tools based on the results from trialling during the first workshop.

5.4.4.1. Improvements from Trialling the Learning Goals Tool.

The design improvements made in the LGT were based on the need to simplify and adapt its taxonomy to suit the needs and expectations of its users. To achieve this, the tool adopted Krathwohl's developments in Bloom's taxonomy, which had already revised and extended the taxonomy by intersecting it with a knowledge dimension. The following design improvements were made:

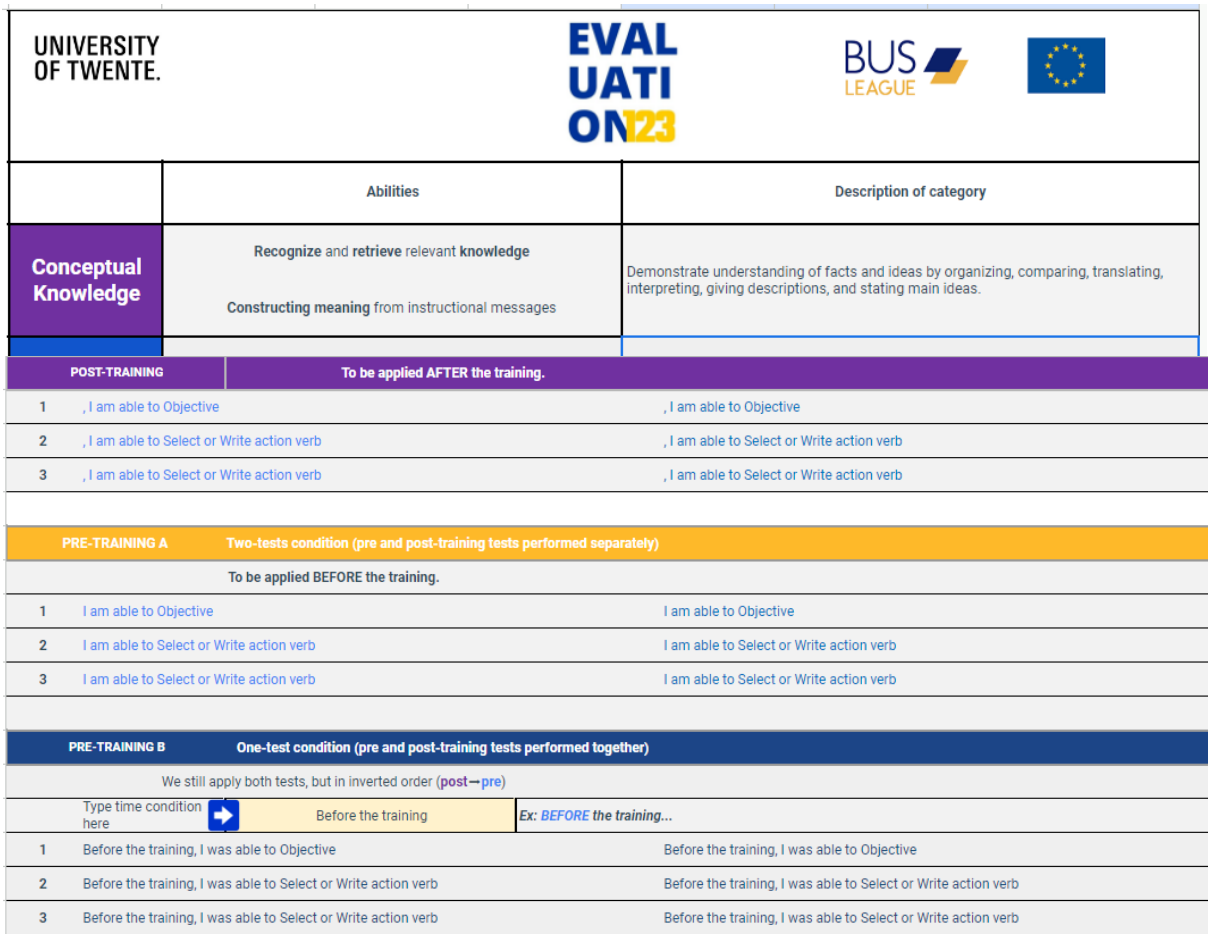
1. Knowledge Dimension: LGT used the Knowledge Dimension instead of cognitive levels as it seemed more appropriate, given that organisations in the construction sector have the perspective of formal learning. The Knowledge Dimension comprises four types of knowledge: Factual, Conceptual, Procedural, and Metacognitive.
    - **Merging Factual and Conceptual Knowledge**: Factual and conceptual knowledge were merged into one, just as it was in the previous version, and named as Conceptual Knowledge.
    - **Renaming Procedural Knowledge**: Procedural Knowledge was renamed as procedural skills to make it more easily recognisable for users.
    - **Renaming Metacognitive Knowledge**: Metacognitive knowledge was renamed as analytical thinking to encompass parts of the high-order cognitive processes.
2. Removing Excessive Visual Clutter: Excessive visual clutter and pop-up information from fields were removed, and instruction is now offered on the screen, next to the corresponding field.
3. Automatization of Process: The WHEN and WHO fields need to be filled in only once, increasing the automatisation of the process.

4. Pretest and Post-test Scenarios: Two different scenarios were created, scenario A (pretest - training - post-test) and scenario B (training - post-test) as the previous pretest and post-test or only post-test scenarios were too confusing. Visualisations were created to understand the occurrences of scenarios.

Overall, the design improvements aimed to simplify and adapt the taxonomy to make it more user-friendly and to improve the overall user experience. The improvements made will help users better understand the purpose and objective of the evaluation, and the taxonomy will be more easily recognisable for users. The automation of the process will make it more efficient, saving time and effort. The new scenarios and visualisations have made it easier for users to understand the occurrences of scenarios, making it more user-friendly (Figure 19).

Figure 19

Learning Goals Tool v.03



5.4.4.2. Improvements from Trialling the Question Packs.

The results from the workshop one did not elicit any major changes to the QP. This might be attributed to the simplicity of the tool, which consists of the digitalisation of survey items in modules.

5.5 A Novel Tool: the Decision Tree (DECISION TREE)

Based on the feedback obtained in the pilots and workshop one, it became apparent that participants still had difficulties visualising the evaluation process. In the design requirements, it had been assumed that providing evaluators with comprehensive information and stepwise guidance would assist them in aligning their expectations and capabilities regarding evaluations. However, it became clear that participants also required assistance in their decision-making processes. To support participants in understanding which decisions have to be made when designing and implementing evaluations, this study created a decision tree (DT).

A DT is a graphical tool that can be used to assist in the decision-making process by breaking down a complex problem or decision into smaller, more manageable parts. It can be especially useful in designing and evaluating programs by helping to define their structure and analyse different scenarios. By depicting decisions into dichotomous decisions (yes/no), which lead to different outcomes, decision trees help users to navigate and reflect on possible scenarios for solving complex problems.

The DT aimed to facilitate the decision-making process by breaking it down into multiple scenarios. The tree also makes it clear to visualise the three different evaluation levels. At each level, the DT presents the decision maker with a set of choices or possible outcomes. Based on these choices or outcomes, the decision decision-maker the best course of action.

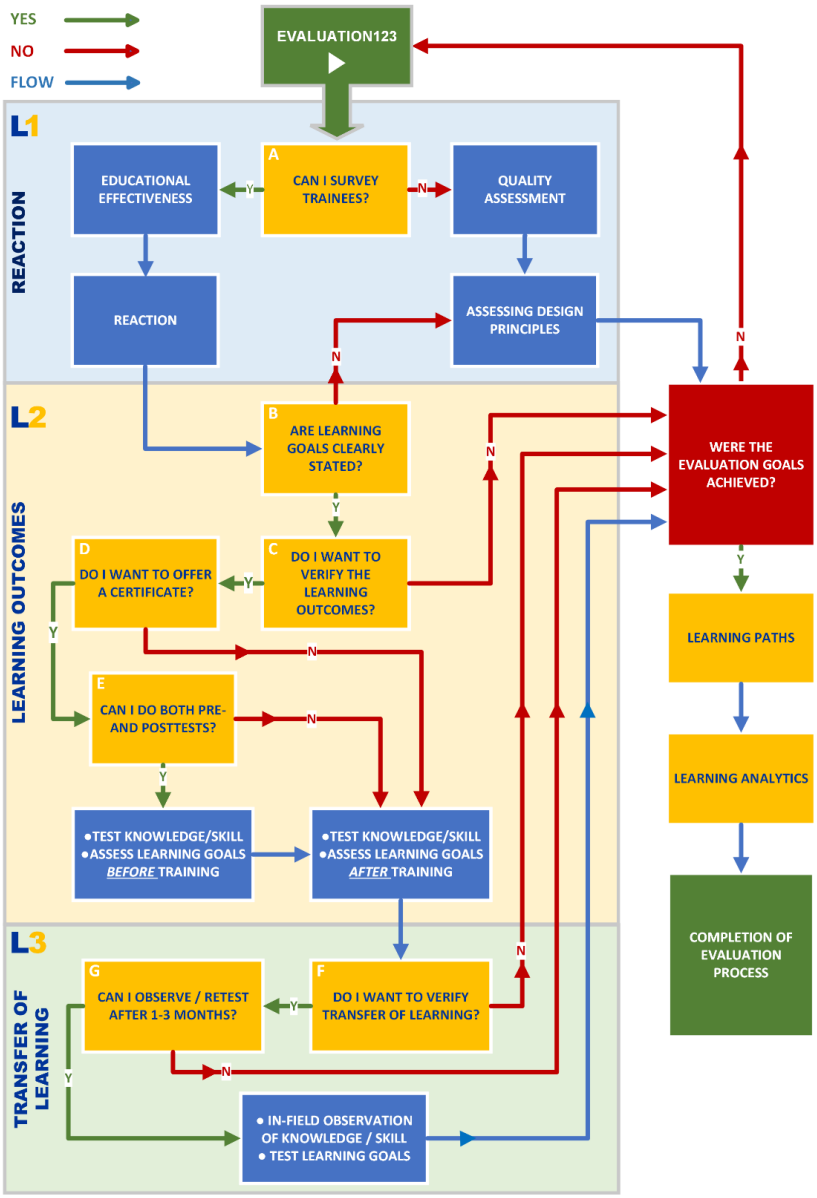
The structure of the DT is designed as a loop, allowing the decision maker to revisit and revise their decisions as new information becomes available or circumstances change.

This is particularly useful when designing and evaluating evaluation programs, as it allows for ongoing reflection and analysis of the program's intentions, level of engagement of resources, and interests.

Each node on the DT directs the decision maker to a respective flashcard containing main considerations or reflections that need to be considered upon a decision. These considerations may include factors such as upskilling goals, stakeholders, available resources, risks and benefits, and potential outcomes (Figure 20).

Figure 20

Decision Tree v.01



The blue boxes in the decision tree represent processes. Yellow boxes represent decision nodes. The red box prompts a reflection on the attainment of evaluation goals and the green box marks the end of the process.

5.6. Workshop 2: Trialling the Decision Tree and the Reference Guide

The primary objective of the second workshop was to present the consortium of partners involved in the BUSLeague project to the revised versions of the preliminary tools, and to trial the DT and the RG as supporting elements for the design of evaluations.

5.6.1. Participants of Workshop 2

The workshop saw participation from nine individuals representing eight different organisations. Prior to the workshop, participants were provided with mock case studies containing the basic information on training program and requirement for evaluation for fictional organisations. All nine participants were appointed representatives in the project for the upskilling activities carried out by their organisations.

5.6.2. Method for Workshop 2

The materials provided to participants included reflective questions aimed at guiding their thought process on the challenges involved in designing evaluations. The overarching goal of the workshop was to empower participants to independently apply the evaluation framework while also testing the updated version of the RG and the DT.

The workshop was conducted via a videoconference and all the sessions were recorded for future reference. The recordings were transcribed using Amberscript. The workshop started with the presentation of the revised Evaluation123 toolset (Reference Guide, LGT and Question Pack) and a brief introduction to the DT. After that, participants were divided into two groups and presented to a possible scenario. The activity consisted of drafting an evaluation plan for the respective scenario using the RG and DT. At the end of

the activity, participants were interviewed, and the utterances were coded according to the SWOT matrix to determine the usability and points of improvement of the tools.

The revised version of the RG brought it from 19 to 46 pages of instructional materials and seventeen pages in the appendix. This increment was largely due to the addition of visual aids and exemplification of procedures and concepts. This version was made available to participants before the workshop, so they could familiarise themselves with the tool and read it beforehand. The DT was added to the RG as an initial guidance before the beginning of the process of designing evaluations at the three levels.

5.6.3. Results of Trialling the Decision Tree and Reference Guide in Workshop 2

5.6.3.1. Results of Trialling the Reference Guide in Workshop 2. The focus group interview at the end of the workshop revealed a positive attitude towards the tool. Participants reported that the RG helped them to clarify their evaluation goals and understand the application of the three levels of evaluation *("I definitely think this could help in the whole process of designing, training and suitable evaluation because it helps in the steps to realize the importance of setting learning goals"*). The language adopted in the RG was considered adequate and not overly technical.

The SWOT analysis revealed that the RG covers a wide range of topics and scenarios and provides a framework for the design of evaluations. It also offers an opportunity for inexperienced evaluators to plan and design evaluations, besides guiding their evaluation efforts. It was mentioned that its applicability could be extended to other projects in energy transition and the construction sector. Yet, the RG may not foresee all the possible scenarios for evaluations (e.g., larger-scale trainings) and may not be easily accessible for evaluators with limited knowledge in English.

5.6.3.2. Results of Trialling the Decision Tree in Workshop 2. All participants were able to accomplish the task within the time proposed and drafted simple evaluation plans. Participants generally had positive feedback about using the DT to help create an

evaluation plan and use the Reference Guide. In several utterances, the DT was referred to as "*very useful*" and "*good tool*" ("*I like the DT…I think this makes it very clear" "very useful" and "I like the DT and I think it is a good way to handle the structure*"). According to the responses obtained, the DT made it clearer to participants the relevance of the definition of learning goals as a core element of evaluation plans. It was also reported that the visualisation of the process afforded by the DT increases understanding of the processes involved and engendered in an evaluation plan. Suggestions were made regarding the placement and layout of a few elements in the DT, such as the detachment of non-level related processes from the main structure.

Based on the SWOT matrix analysis, the DT has several strengths, including its ability to help users visualise processes and make design decisions based on capabilities and intentions. Additionally, the tool seems easy to navigate and does not require much previous explanation. The tool also presents opportunities for users to explore possibilities and plan evaluation programs. However, there are also weaknesses to consider, specifically the presence of green and red boxes inside the plan that can lead to wrong interpretations, as they may seem to belong to a certain level. This lack of clarity may also be a threat, as it may make the process appear too complex to users.

Overall, the tool has significant potential as a planning and evaluation program tool, but it may require further development to address the weaknesses and provide users with a clearer understanding of the process.

5.7. Design improvements resulting from Workshop 2

5.7.1. Improvements from Trialling the Reference Guide

The trialling and feedback from users demonstrated that the RG had reached a near-maturity state, requiring only minor changes. However, the researcher in this study has identified a few points of improvement that could increase the readability and capability of the RG in conveying more precise and easy-to-follow instructions. This included breaking down

instructions into smaller steps to facilitate their reading and following. Some chapters and items were colour-coded, to indicate they belong to the same procedure or process. This also reflected in the layout of the RG, which should reflect principles of instructional and multimedia design (Mayer, 2001).

5.7.2. Improvements from Trialling the Decision Tree

This report outlines the improvements that were made to the designed tool (DT) based on feedback received during the debriefing sessions. The DT was designed to facilitate a complex process, and the improvements aimed at enhancing the user experience, reducing confusion, and increasing efficiency.

1. Highlighted start of the process: One of the improvements made to the DT was highlighting the start of the process. This was done to ensure that users understood where to begin and could easily locate the starting point. By doing so, it eliminated any confusion and improved the user experience.
2. Flipped sides: Another improvement was flipping the sides from left to right reading. This was done to align the DT with the natural reading direction of the users. This simple change made a significant difference in the user experience, and feedback received indicated that users found it easier to follow the process.
3. Removed extraneous processes: During the debriefing sessions, feedback revealed that some processes were unnecessary and could be removed. As a result, extraneous processes (*achievement of learning goals, learning paths, analytics and completion)* were eliminated from the level plane, streamlining the DT, and making it more user-friendly.
4. Quality assessment as a parallel route: In the initial design, quality assessment was considered an alternate route. However, feedback indicated that it was a parallel route, and as such, it was made more prominent and included as part of the primary process. This adjustment improved the efficiency of the process and eliminated confusion.

5. Removed page numbers from nodes: Another improvement was removing page numbers from nodes. Instead, they were included in flashcards. This change reduced visual clutter, making the DT less overwhelming and easier to follow.
6. Reinforced visual elements: Finally, visual elements were reinforced, making the DT more aesthetically pleasing and user-friendly. This included the use of colour-coding, shapes, and icons. The reinforced visual elements improved the user experience and made the DT more accessible to a wider range of users.

Overall, the improvements made to the designed tool addressed feedback received during the debriefing sessions, resulting in a more user-friendly and efficient tool. The highlighted start of the process, flipped sides, removed extraneous processes, quality assessment as a parallel route, removed page numbers from nodes, and reinforced visual elements all played an essential role in making the DT more accessible and easier to use.

6. Evaluation and Reflection

McKenney and Reeves (2018) distinguish three main types of tests in the evaluation phase, with different foci: *Alpha (internal structure), Beta (use in context) and Gamma (effects).* This study carried out a Beta-test on the Evaluation123 by running a try-out of the toolset (RG, LGT, QP and DT) in a relevant setting. The beta-test aimed to gain an understanding on the functionality and interaction of the Evaluation123 in context. Furthermore, a reflection process contributed to a deeper understanding of the Evaluation123, its intended purpose, and the impact it generated.

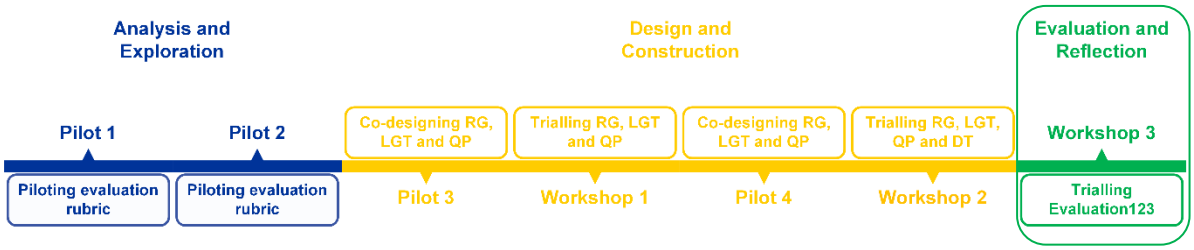
To guide the assessment of usability and perceived usefulness of the designed tools, the following research question was posed:

RQ.3: *How do these designs perform in guiding the design of evaluation programs in organisations in the construction and energy efficiency sector?*

An overview of the activities carried out in this study can be seen in Figure 21.

Figure 21

Overview of Activities Carried Out in the Design and Construction Phase



6.1. Participants in Piloting of Design of Evaluations

The beta-testing of the Evaluation123 took place in a face-to-face event held during a general consortium meeting with representatives of all partners in BUSLeague, with a total of fourteen participants from the nine organisations involved in the project. A total of eleven participants completed the survey.

6.2. Methodology and Instrumentation of the Evaluation and Reflection Phase

In previous instances of piloting and workshops, participants were presented with case studies, meant to facilitate, and guide the activities performed during those events. However, to test the Evaluation123 in a realistic context, participants were requested to prepare for the beta-testing workshop by selecting and bringing along data about one or more of their upcoming trainings. This data included the title of the training, expected number of participants and the goals of the training. During the workshop, the fully developed version of the Evaluation123 toolset was presented. After a brief introduction to the activity, participants were asked to draft an evaluation plan for their trainings using the proposed tools. The activity was guided with the use of a checklist form, where they could note down the results of their design.

The beta-testing of the toolset aimed at assessing the usability and perceived usefulness of the proposed solution. The drafted evaluation designs were collected for analysis and feedback. After the activity, participants were asked to perform two assessment activities: an evaluation survey of the toolset, and to complete a plus and minus test on the Reference Guide.

The survey aimed to collect quantitative data on the usability and feasibility of the Reference Guide and associated tools as guiding instruments for designing evaluation plans. It comprised two elements, a questionnaire, and a comprehension test. The questionnaire was inspired in the UTAUT model (Dillon, 2006) and consisted of 17 questions on a 6-level agreement scale (Likert), from *totally agree* to *totally disagree (see Appendix F)*. The first five questions related to the facilitating conditions of the RG (e.g., *it is easy to look up things in the Reference Guide)*. Questions 6-10 measure attitude towards using the RG (e.g., *the Reference Guide will be useful for planning evaluations in different contexts*). Questions 10-11 referred specifically to the facilitating conditions of the DT (e.g., *the Decision Tree in the Reference Guide facilitates planning an evaluation*). The LGT was assessed in questions 12-

14 (e.g., *the Learning Goals Tool helps me specify the ULOs to units that can be evaluated*). Questions 15-16 assessed the QPs (e.g., *the Questionnaire Packs help me select items for my surveys*). Lastly, question 17 was not meant to assess the Evaluation123, but aimed at assessing the inclusion of the checklist used during the beta-testing workshop as an additional auxiliary tool in the appendix of the RG.

The plus (+) and minus (-) test is commonly used for subjectively evaluating graphic and written materials. The test consists of asking participants to assign the corresponding marks to elements in a material, for a range of criteria, such as comprehensibility, appreciation, or relevance (de Jong & Schellens, 1999). Despite having its validity compromised by the self-reported nature of the test, the plus and minus test has been reported to be a useful instrument for identifying problems (de Jong & Rijnks, 2006), collecting feedback, and informing design improvements (de Jong & Schellens, 1999). By employing the plus-minus technique, one can explore participants' perspectives on the document's clarity and identify their emotional responses to its content. Explicitly positive comments or markings were designated as positive, while negative markings and explicit criticisms were designated as negatives. Neutral messages encompassed questions, suggestions, and corrections of typographical errors.

Based on a recommendation from (De Jong & Schellens, 2000), a brief five-question multiple-choice comprehension test was coupled with the survey. The addition of comprehension tests alongside the plus and minus method serves to provide a more comprehensive evaluation of readers' comprehension abilities. These tests typically consist of questions that require readers to recall specific details, make inferences, or demonstrate a deeper understanding of the text. The five questions abridged some of the main concepts contained in the Evaluation123 framework and consisted of questions such as “*In the Evaluation123, the levels are independent of each other*” and “*Knowledge and skill tests are the only valid ways of assessing learning outcomes*”. All questions were true or false, and a “I do not know” option was used to minimise potential deviations due to guessing.

6.3. Data Analysis of the Evaluation and Reflection Phase

6.3.1. *Data Analysis for the Survey and Comprehension Test*

The items in the questionnaire survey were organised in five distinct clusters, measuring different constructs. All items were measured in a six-point Likert agreement scale (totally disagree – totally agree). The scale was coded 1 (totally disagree) to 6 (totally agree) and the measures of central tendency (means, median, modes and standard deviations) were calculated using statistical software (SPSS). This allowed a better understanding of the data, including the visualisation of their distribution.

The main findings of the survey and the detailed results can be found in Appendix G.

6.3.2. *Data Analysis for the Plus and Minus Test*

The purpose of adopting the plus and minus test in this study was to gain a deeper understanding of the quality of the RG from the perspective of the readers. However, this type of testing requires participants to rate every paragraph and item on a written material. Due to limitations in time during the try-out workshop, participants could not read the entire RG in the proposed time (30 minutes). Therefore, they were asked to skim-read and rate every item they found relevant, either with a + or a – sign. As not all participants rated the same items, the test did not yield consistent results that could allow for inferences on the quality of the RG but could still serve as indicators for points of improvement and values of the guide.

6.4. Results of the Evaluation and Reflection Phase

6.4.1. *Results of the Survey and Comprehension Test*

The results indicated that the majority of the BUSLeague consortium partners found the reference guide, decision tree, learning goals tool, and question packs to be usable and useful for conducting evaluations of their trainings (see Table 4). There were no disagreements with any of the statements regarding usability and usefulness. However, there

is still room for improvement in bridging the learning goals tool with existing ULOs and enhancing the comprehensibility of the reference guide. Below are the main insights obtained from the analysis of the survey data.

Table 4

Key Statistics from the Survey

	Valid N	Mean	Median	Mode	Std. Deviation
Usability RG	11	4.62	4.40	4.40	0.540
Usefulness RG	11	5.00	5.00	5.00	0.798
Usefulness DT	11	5.18	5.00	6.00	0.751
Usefulness LGT	11	4.52	4.67	4.00	0.848
Usefulness QP	11	5.14	5.00	5.50	0.393
Usefulness Checklist	11	4.91	5.00	5.00	0.701

- **Usability of the reference guide**: respondents generally agreed that the RG is easy to read and navigate (*M = 4.62, SD = 0.540*), and that it also facilitates comprehension of the topics. Nonetheless, respondents also somewhat agreed that the RG uses too many technical words.
- **Usefulness of the reference guide**: Respondents agreed that the reference guide helps in comprehending evaluation needs and possibilities (*M = 5.0, SD = 0.798*). It provides a comprehensive overview and is valuable for planning evaluations in various contexts. They also felt confident in using the reference guide to plan evaluations.
- **Usefulness of the decision tree**: Respondents agreed that the decision tree assists in planning evaluations and maximising outcomes within their limitations by weighing evaluation benefits and efforts (*M = 5.18, SD = 0.751*).
- **Usefulness of the learning goals tool**: Respondents agreed that the learning goals tool supports the design of their evaluations (*M = 4.52, SD = 0.848*). They somewhat agreed

that it helps in specifying ULOs (Ultimate Learning Objectives) and clarifying learning objectives.

- **Usefulness of the question packs**: Respondents agreed that the question packs aid in selecting survey items (*M = 5.14, SD = 0.393*). They strongly agreed that the digital format in Google Forms enhances their usability.

The results from the comprehension test suggested by De Jong & Schellens (2000) revealed that participants in general performed well. When considering all three possible answers (right, wrong, and I do not know), the data shows that 63.6% of questions were answered correctly, 21.8% were answered wrongly, and 14.8% were marked as "I do not know." This indicates that most respondents were able to provide the correct answers, while a significant portion admitted uncertainty by selecting the "I do not know" option. To focus specifically on valid answers (right and wrong), we find that 73.8% of responses were correct, while 26.2% were incorrect.

Upon examining individual user performance, we observe that five users achieved a score of 80% on the test, whereas four users scored 60%. The analysis of the multiple-choice question responses reveals that, overall, respondents performed reasonably well, with most questions being answered correctly.

6.4.2. *Results of the Plus and Minus Test*

The ratings aided in identifying patterns and a total of 62 trends within the responses. In total, sixty-two sections were categorised as positive, twenty-two sections as negative and nine sections received neutral messages. Subsequently, a condensed overview of the significant insights derived from these ratings will be provided.

- **Appreciation for Introduction to Evaluations and Questioning Techniques** - The introduction to evaluations and the section about making the right questions received positive appraisals. Respondents admired the use of simple and straightforward language, suggesting that the style and language were appropriate and not overly

complex. This indicates that the goals and aims of the evaluation aligned with the expectations of representatives from the construction sector.

- **Emphasis on Visual Elements**: In general, respondents appreciated the presence of figures in the text. This positive reception aligns with multimedia design principles that suggest imagery improves comprehension. The use of visual aids could prove beneficial in future evaluations, enhancing participants' understanding and engagement.
- **Reflection on Resource Availability** - The reflection on the availability of resources for evaluation garnered positive feedback. Participants found the elements presented for consideration relevant to the construction industry, specifically highlighting time, human resources, and materials. This demonstrates the importance of addressing these factors within evaluations to ensure practicality and applicability.
- **Mixed Reactions to the use of Validated Measurement Instruments** - The presentation of the validated measurement instruments from the initial rubric and Question Packs received mixed reactions. While some respondents marked it positively, others expressed concerns over vague and unclear definitions. This disparity might stem from the resistance organisations still exhibit toward the validity of reaction questions, particularly when employing complex, multi-item instruments. Many organisations prefer single-item questions to measure constructs, rather than utilizing multiple items.
- **Positive Reception of Learning Levels and SMART LG Definition** - Learning levels within the test were favourably appraised by respondents, indicating their perceived value in evaluating performance. The definition of SMART LG, which represents specific, measurable, attainable, relevant, and time-bound learning goals, received praise as a concise and effective summary.
- **LGT: Consolidation and Value** - Multiple positive markings were observed regarding the Learning Goal Tool (LGT). This suggests a greater appreciation for LGT compared to single positive markings. Respondents indicated that LGT appeared to be consolidated

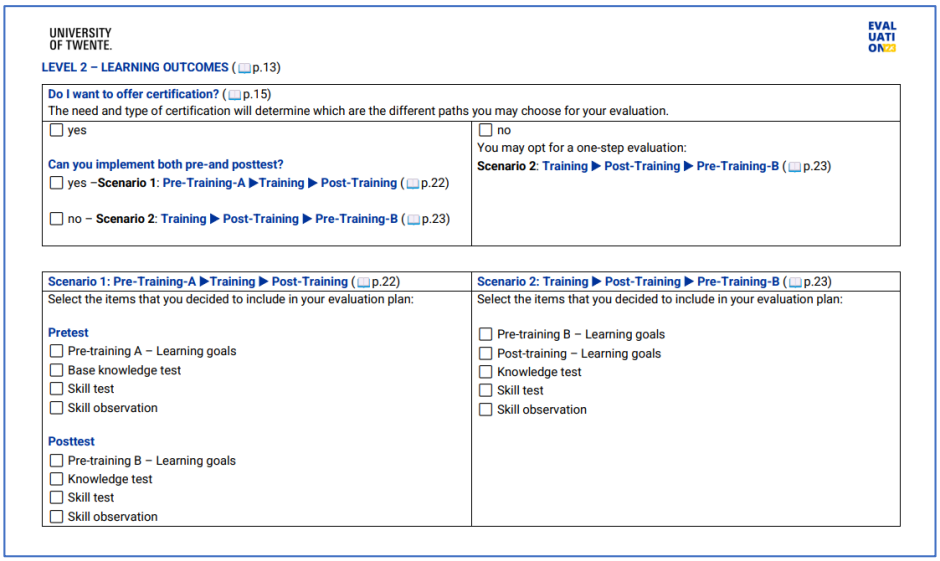
and valued at this stage. They also recognised LG as the focal point for both training design and evaluation efforts, highlighting its significance in the overall process.

6.4.3. *Checklist*

The checklist was conceived as a worksheet to assist evaluators to structure and design their evaluation plans during the workshop. It provided a form-like structure in which participants could record and trace all the most relevant items when designing an evaluation plan (see Figure 22). It was structured according to the three levels of evaluation proposed by the Evaluation123 framework and brought the most important tools, such as the LGT and the DT (see Appendix H). The checklist also contained the DT, and participants were instructed to mark down their decision routes.

Figure 22

Sample Page of the Checklist Used in the Try-Out Workshop.


UNIVERSITY OF TWENTE.

EVAL UATI ON

LEVEL 2 – LEARNING OUTCOMES (p.13)

Do I want to offer certification? (p.15)
The need and type of certification will determine which are the different paths you may choose for your evaluation.

☐ yes

Can you implement both pre-and posttest?
☐ yes –**Scenario 1**: **Pre-Training-A ▶Training ▶ Post-Training** (p.22)

☐ no – **Scenario 2**: **Training ▶ Post-Training ▶ Pre-Training-B** (p.23)

☐ no
You may opt for a one-step evaluation:
Scenario 2: **Training ▶ Post-Training ▶ Pre-Training-B** (p.23)

Scenario 1: Pre-Training-A ▶Training ▶ Post-Training (p.22)
Select the items that you decided to include in your evaluation plan:

Pretest
☐ Pre-training A – Learning goals
☐ Base knowledge test
☐ Skill test
☐ Skill observation

Posttest
☐ Pre-training B – Learning goals
☐ Knowledge test
☐ Skill test
☐ Skill observation

Scenario 2: Training ▶ Post-Training ▶ Pre-Training-B (p.23)
Select the items that you decided to include in your evaluation plan:

☐ Pre-training B – Learning goals
☐ Post-training – Learning goals
☐ Knowledge test
☐ Skill test
☐ Skill observation

The last question in the survey questionnaire aimed at assessing the potential of including the checklist as an additional tool to the Evaluation123 toolset. Participants agreed (81.8%) that the checklist was useful for planning an evaluation (*M=4.91, SD = 0.701*).

6.4.3.1. Insights from the Checklist.

The checklist was intended to guide the activities during the try-out workshop, and enable participants to track their decisions and choices during the design of the evaluation plans for their trainings. After the workshop, these checklists were collected to provide feedback for participants on their evaluation plans and suggest possible improvements. The evaluation plans drafted in the checklists were analysed and resulted in ten individual reports. The reports provided feedback on the alignment of evaluation plans with the upskilling programs, reflected on their decisions and suggested corrections for conflicting interests in the plans.

This action resulted in a secondary source of data, from which this study could gain a greater understanding of the Evaluation123 and its associated tools. The evaluation plans drafted during the workshop were analysed and feedback was provided to the users. In total, ten evaluations plans were documented. All the plans analysed reached at least a level 2 evaluation, which means they foresaw the assessment of learning outcomes. Only four of them aimed at level 3 – transfer of learning, which implies the observation and testing of knowledge and skills over a period of time. In general, the defined LG defined during the workshop were sufficiently clear and complied with the SMART LG directives. Participants were asked to mark their decision paths in the DT, and the results indicated that the paths marked were congruent with the plans drafted.

6.4.4. *Integration of Results of the Evaluation and Reflection Phase*

The outcomes of the Evaluation and Exploration phase provided valuable insights into the values and shortcomings of the developed Evaluation123 framework and its tools. The results of the evaluation phase indicate that the reference guide, decision tree, learning goals tool, and question packs were found to be usable and useful by most participants.

The different methods for data collection and analyses enabled this study to gain a wider perspective on the values and shortcomings of the Evaluation123 and its toolset. Whereas the survey measured an overall positive reaction and perceived usefulness for the Evaluation123, the plus and minus test demonstrated that the efforts in making evaluation

processes less complex were fruitful. The analysis of the evaluation plans designed during the try-out workshop also revealed that the overall quality of the products was satisfactory, and that participants were able to plan more elaborate evaluation plans.

However, there is room for improvement in terms enhancing the comprehensibility of the reference guide. The plus and minus test, which asked participants to provide feedback on the materials, revealed positive responses to the introduction, visual elements, reflection on resource availability, learning levels, and the definition of SMART learning goals. Mixed reactions were observed for the rubric and Question Pack items. Additionally, the evaluation plans produced during the workshop were analysed. All plans reached at least level 2 evaluation, assessing learning outcomes, and some aimed at level 3 evaluation, which involves observing and testing knowledge and skills over time. This might demonstrate the capability of the designed tools in fostering higher levels of evaluations. The learning goals defined during the workshop were generally clear and aligned with the SMART LG directives, and the decision paths marked on the decision tree were congruent with the plans. Overall, the Evaluation and Reflection phase provided insights into the usability and usefulness of the tools, as well as the strength of the evaluation plans created using them.

The opportunity of implementing a comprehensive evaluation process during the workshop was restricted by the availability of time. The workshop took place during a consortium meeting, in which several topics related to the BUSLeague project were discussed. The plus and minus test, for example, was impacted by these restrictions, as it commonly requires that respondents read and rate each item in the material. Additionally, the time restriction also affected the ability of less experienced participants to design consistent evaluation plans.

6.5. Conclusion

The conclusion chapter of this educational design research study serves as a final synthesis and reflection of the research findings and their implications. It begins by summarizing the main findings and key insights obtained through the research process. The

chapter then delves into the broader implications of these findings, discussing their significance and relevance to the field of study and potential practical applications. Additionally, it addresses any limitations or gaps in the research and suggests areas for further exploration.

6.5.1. *Discussion*

The aim of this study was to determine how to guide the evaluation of effectiveness of upskilling micro-trainings in energy efficiency for organisations in the construction and energy sectors, with limited (human, financial, material) resources and expertise. This was accomplished by using DBR as a methodology, in a process of co-designing the solution with other stakeholders in the BUSLeague project.

In the Analysis and Exploration phase of the study, the capabilities, limitations and needs of the stakeholders in designing evaluation plans were scrutinised, as well as the hinderances in the use of the evaluation rubric proposed at the BUSLeague project. The analysis of documents pertaining the project revealed that the evaluation of upskilling activities was deemed by the project consortium as a key factor for success, as it directly affects the recognition of professional skills across borders within the EU. Interviews with stakeholders revealed that evaluation plans were mostly done experimentally, meaning the designed plans were not supported by any tested or valid methodology for evaluation. This lack of experience and knowledge by evaluators reflected directly on the definition of learning goals, which usually led to overly broad, unmeasurable objectives. Ultimately, the unfamiliarity with evaluation procedures and design resulted in a misalignment between their expectations and real capabilities of measuring learning outcomes. This lack of understanding reflected on their comprehension of the existing evaluation rubric and impacted its adoption and usability within BUSLeague. Thus, it became clear that the existing rubric would require further development, also implying the development of tools that would support its implementation.

To cope with these challenges, in the Design and Construction phase, prototypes for diverse tools were co-designed and tested. This study proposed the creation of a Reference Guide, as an instrument to scaffold evaluators in their design of evaluation plans. The co-designing process informed researched that any solution proposed would have to adopt a very simple, easy to follow language and communication, since evaluation plans may need to be drafted by unexperienced evaluators. It is customary for the industry that trainers may also be responsible for the evaluation. However, these trainers are sometimes specialised craftsman and technicians with no experience in education. This yielded the need for an illustrative and concise guidance, with colloquial language and exemplified instructions. Given the perceived complexity of a few processes, the reference guide was complemented by other tools. To cope with the difficulty in visualizing the evaluation process, its requirements and the decisions that must be considered upon its execution, a decision tree was created. The DT was reported to facilitate the process by allowing users to understand all the steps involved in designing an evaluation plan and providing a visual roadmap from which they could envision possible tracks and align their expectations with regards to outcomes.

A digital Learning Goals Tool was developed to assist users in defining clear, SMART learning goals. The tool was designed to be intuitive, contain all the required elements in a LG and prevent the creation of unmeasurable learning objectives. This was achieved by simplifying well-known taxonomies and translating them into more accessible terms. The collection of valid measuring instruments contained in the rubric was rearranged and their adoption was facilitated by their digitalisation. By offering the survey instruments in easy-to-import digital packs, it was expected to foster their use and mitigate their eventual dismemberment, which would compromise their reliability and validity.

The comprehensive support offered by the EVALUATION123 framework, centred around the reference guide, has been well-received by the BUSLeague consortium partners. This step-by-step guidance has proven invaluable, as evidenced by its adoption by related

EU projects, such as BUS Go Circular and the 4D Digitalization of Practice Placements, even during the project's runtime. Partners within the consortium have expressed their appreciation for EVALUATION123, recognizing its potential to enhance their work in providing training for construction and building professionals.

6.5.2. *Implications for Practice and Theory*

By addressing requirements and attending to needs from a diverse group of stakeholders, this study has been able to develop a framework and tools that could be easily adopted by different settings and organisations in the construction sector and beyond. A harmonised and reliable structure is essential in the efforts of the industry to bring mutual and cross-border recognition of skills throughout the european market. Secondly, by replacing experimental practices by a scientifically oriented model for evaluations, organisations in the construction sector will get more reliable measurements and assessments of their training and upskilling efforts.

The Evaluation123 also focused on the development of resources and tools for evaluations in the construction sector, which is not usually the focus of scientific studies. Studies dedicated to evaluation and assessment are usually set in academic environments or in settings where high-level cognitive skills are expected. Thus, this study hopes to have contributed to scientific knowledge on the evaluation of procedural skills.

6.5.3. *Limitations and Recommendations*

Given the scope of the project under which this study was conducted, a few limitations and restrictions applied. Organisations usually have limited human, material, and time resources for evaluations, and set their focus on the training element. This limits their interest and pushes into the adoption of simpler means of evaluation. None of the participating organisations affirmed having the knowledge to perform statistical analysis, and all the data collected through evaluations were only descriptively analysed. However, given the high quality of data that the digitalisation of evaluation processes can afford, future

studies should concentrate on the analysis of these data and use of learning analytics to enrich scientifically knowledge on education in non-academic settings. Additionally, the co-designing and immediate implementation of the tools – given the need to develop a comprehensive toolset and evaluation framework– did not allow this study to analyse them in depth.

Furthermore, studies are needed on the revision of taxonomies to better accommodate procedural skills. Despite Bloom's being the most well-known taxonomy, corroborate by the acclaimed revision made by Krathwohl, both are targeted at classifying the cognitive processes and levels of complexity. Other taxonomies such as SOLO (*Structured Observed Learning Outcome*) (Biggs & Collis, 1982) and the *Psychomotor Domain* (Simpson, 1971) focus on cognitive understanding and depth of knowledge, and on the development and assessment of physical skills and abilities, respectively. However, none of them seem to fit appropriately the categorisation of practical and procedural skills developed at the workplace, at least, not to a point that they are practical or accessible for non-academic purposes.

7 References

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. Journal of the Medical Library Association : JMLA, 103(3), 152–153. https://doi.org/10.3163/1536-5050.103.3.010

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). a Meta-Analysis of the Relations Among Training Criteria. Personnel Psychology, 50(2), 341–358. https://doi.org/10.1111/j.1744-6570.1997.tb00911.x

Alvares, O. (2003). Goals and Objectives. Critical Reviews in Oral Biology and Medicine, 14(1), 1. https://doi.org/10.1177/154411130301400101

American evaluation association. (2012). Guiding principles. The UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions. https://doi.org/10.1007/978-3-642-25995-1_4

Anaya, G. (1999). College impact on student learning: Comparing the use of self-reported gains, standardised test scores, and college grades. Research in Higher Education, 40(5), 499–526. https://doi.org/10.1023/A:1018744326915

Anderson, D. ., & Krathwol, L. . (2001). REVISED Bloom ' s Taxonomy Action Verbs: A taxonomy for learning, teaching, and assessing, Abridged Edition. 2001. www.apu.edu/live.../blooms_taxonomy_action_verbs.pd

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? Educational Researcher, 41(1), 16–25. https://doi.org/10.3102/0013189X11428813

Arico, F., Gillespie, H., Lancaster, S., Ward, N., & Ylonen, A. (2018). Lessons in learning gain: Insights from a pilot project. Higher Education Pedagogies, 3(1), 249–265. https://doi.org/10.1080/23752696.2018.1454845

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & Von Davier, A. A. (2019). The expanded Evidence-Centered-Design (e-ECD) for learning and assessment systems: A framework to incorporating learning goals and processes within assessment design. Frontiers in Psychology, 10(APR). https://doi.org/10.3389/fpsyg.2019.00853

Armstrong, P. (2010). Bloom's Taxonomy. Vanderbilt University Center for Teaching. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Athiyaman, A. (1997). Linking student satisfaction and service quality perceptions: the case of university education. European Journal of Marketing, 31(7), 528–540. https://doi.org/10.1108/03090569710176655

Axtell, C. M., Maitlis, S., & Yearta, S. K. (1997). Predicting immediate and longer-term transfer of training. Personnel Review, 26(3), 201–213. https://doi.org/10.1108/00483489710161413

Bakker, E. J. M., Verhaegh, K. J., Kox, J. H. A. M., van der Beek, A. J., Boot, C. R. L., Roelofs, P. D. D. M., & Francke, A. L. (2019). Late dropout from nursing education:

An interview study of nursing students' experiences and reasons. Nurse Education in Practice, 39(June), 17–25. https://doi.org/10.1016/j.nepr.2019.07.005

Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. Evaluation and Program Planning, 27(3), 341–347. https://doi.org/10.1016/j.evalprogplan.2004.04.011

Berendt, B., Vuorikari, R., Littlejohn, A., & Margaryan, A. (2012). LEARNING ANALYTICS AND THEIR APPLICATION IN TECHNOLOGY-ENHANCED PROFESSIONAL LEARNING (Issue 4).

Biggs, J. (1999). Teaching for Quality Learning at University. Assessing for learning quality: II. Practice. Teaching for Quality Learning at University, September, 165–203.

Bluestone, J., Johnson, P., Fullerton, J., Carr, C., Alderman, J., & BonTempo, J. (2013). Effective in-service training design and delivery: Evidence from an integrative literature review. Human Resources for Health, 11(1). https://doi.org/10.1186/1478-4491-11-51

Božana Knežević. (n.d.). Revisiting the learning outcomes debate in the context of student self- assessment of receptive and productive skills. 1–14.

Cahapay, M. (2021). Kirkpatrick Model: Its Limitations as Used in Higher Education Evaluation. International Journal of Assessment Tools in Education, 8(1), 135–144. https://doi.org/10.21449/ijate.856143

Campbell, R., Okey, J., Quitadamo, I. J., Kurtz, M. J., Paul, R. and Nosich, G., & Rathburn, B. S. (1993). Curriculum & Leadership Journal _ Skills for the 21st Century_ teaching higher-order thinking. CBE - Life Sciences Education, 6, 1–15. https://doi.org/10.1187/cbe.06

Cant, R., & Levett-Jones, T. (2021). Umbrella review: Impact of registered nurses' continuing professional education informed by contemporary reviews of literature. Nurse Education in Practice, 50(November 2020), 102945. https://doi.org/10.1016/j.nepr.2020.102945

CEDEFOP. (2020). European qualifications framework. Initial vocational education and training: focus on qualifications at levels 3 and 4. Publications Office of the European Union. https://doi.org/10.2801/114528

Chao, R. J., & Chen, Y. H. (2009). Evaluation of the criteria and effectiveness of distance e-learning with consistent fuzzy preference relations. Expert Systems with Applications, 36(7), 10657–10662. https://doi.org/10.1016/j.eswa.2009.02.047

Chauhan, R., Ghosh, P., Rai, A., & Kapoor, S. (2017). Improving transfer of training with transfer design. Journal of Workplace Learning, 29(4), 268–285. https://doi.org/10.1108/jwl-08-2016-0079

Cheng, E. W. L., & Hampson, I. (2008). Transfer of training: A review and new insights. International Journal of Management Reviews, 10(4), 327–341. https://doi.org/10.1111/j.1468-2370.2007.00230.x

Cunningham, G. B. (2007). Development of the Physical Activity Class Satisfaction Questionnaire (PACSQ). Measurement in Physical Education and Exercise Science, 11(3), 161–176. https://doi.org/10.1080/10913670701326443

Czaja, R., & Blair, J. (2011). Questionnaire Design: Testing the Questions. Designing Surveys, 103–123. https://doi.org/10.4135/9781412983877.n6

Dauenbeimer, D. G., Stablberg, D., Spreemann, S., & Sedikides, C. (2002). Assessmen0The Intricate Role ofTrait Modifiability in the Self-Evaluation Process. Psychologie Schweizerische Zeitschrift Für Psychologie Und Ihre Andwendungen, 15, 89–112.

De Jong, M., & Rijnks, D. (2006). Dynamics of iterative reader feedback: An analysis of two successive plus-minus evaluation studies. *Journal of Business and Technical Communication*, *20*(2), 159–176. https://doi.org/10.1177/1050651905284401

de Jong, M., & Schellens, P. J. (2000). Toward a document evaluation methodology: What does research tell us about the validity and reliability of evaluation methods? IEEE Transactions on Professional Communication, 43(3), 242–260. https://doi.org/10.1109/47.867941

de Laat, M., & Schreurs, B. (2013). Visualising Informal Professional Development Networks: Building a Case for Learning Analytics in the Workplace. American Behavioral Scientist, 57(10), 1421–1438. https://doi.org/10.1177/0002764213479364

de Long, M., Winter, D., & Yackel, C. A. (2005). Student learning objectives and mathematics teaching. Primus, 15(3), 226–258. https://doi.org/10.1080/10511970508984119

Diamantidis, A. D., & Chatzoglou, P. D. (2014). Employee post-training behaviour and performance: Evaluating the results of the training process. International Journal of Training and Development, 18(3), 149–170. https://doi.org/10.1111/ijtd.12034

Doktor, F. (2019). Public document-Database of common descriptors for defining required skills, knowledge, and competences of the targeted construction professionals.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology, 41(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

E. A. Ruona, W., Leimbach, M., F. Holton III, E., & Bates, R. (2002). The relationship between learner utility reactions and predicted learning transfer among trainees. International Journal of Training and Development, 6(4), 218–228. https://doi.org/10.1111/1468-2419.00160

European Commission. (2011). USING LEARNING OUTCOMES. In European Qualifications Framework series: note 4. Publications Office of the European Union. https://doi.org/102766/17497

Ferguson, L. M. (1998). Writing learning objectives. Journal of Nursing Staff Development : JNSD, 14(2), 87–94. http://www.ncbi.nlm.nih.gov/pubmed/9592434

Fessl, A., Maitz, K., Dennerlein, S., & Pammer-Schindler, V. (2021). The Impact of Explicating Learning Goals on Teaching and Learning in Higher Education:

Evaluating a Learning Goal Visualisation. In T. De Laet, R. Klemke, C. Alario-Hoyos, I. Hilliger, & A. Ortega-Arranz (Eds.), Lecture Notes in Computer Science (Vol. 12884, pp. 1–15). Springer International Publishing. https://doi.org/10.1007/978-3-030-86436-1_1

Filliettaz, L. (2014). Understanding Learning for Work: Contributions from Discourse and Interaction Analysis. https://doi.org/10.1007/978-94-017-8902-8_9

Forehand, M. (2005). Bloom's Taxonomy. In Bloom's Taxonomy: Original and Revised. https://doi.org/10.4135/9781412957403.n51

Francis, L. J., & Smith, B. J. (2015). Toward Best Practice in Evaluation: A Study of Australian Health Promotion Agencies. Health Promotion Practice, 16(5), 715–723. https://doi.org/10.1177/1524839915572574

Frechtling, J. A. (2015). Logic Models. In International Encyclopedia of the Social & Behavioral Sciences: Second Edition (Second Edi, Vol. 14). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.10549-5

Friedrich, J. (1996). Assessing Students' Perceptions of Psychology as a Science: Validation of a Self-Report Measure. Teaching of Psychology, 23(1), 6–13. https://doi.org/10.1207/s15328023top2301_1

Frye, A. W., & Hemmer, P. A. (2012). Program evaluation models and related theories: AMEE Guide No. 67. Medical Teacher, 34(5). https://doi.org/10.3109/0142159X.2012.668637

Gao, Y., González, V. A., & Yiu, T. W. (2017). Serious Games vs. Traditional Tools in Construction Safety Training: A Review. Lean and Computing in Construction Congress - Volume 1: Proceedings of the Joint Conference on Computing in Construction, 6(July), 653–660. https://doi.org/10.24928/JC3-2017/0070

Gary, R., Pike, G. R., & Hall, T. (1994). TITLE ACHIEVEMENT-TEST SCORES.

George T. Doran. (1981, November). There's a S.M.A.R.T. way to write management's goals and objectives. AMA FORUM. https://community.mis.temple.edu/mis0855002fall2015/files/2015/10/S.M.A.R.T-Way-Management-Review.pdf

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. New Directions for Institutional Research, 2005(127), 73–89. https://doi.org/10.1002/ir.156

Hirose, L., Dennerlein, S. M., Jong, L. De, & Endedijk, M. D. (2022). Guiding Evaluation of Trainings in the Field of Energy Transition. 30. https://doi.org/10.31235/osf.io/yb8w5

Jadallah, H., Friedland, C. J., Nahmens, I., Pecquet, C., Berryman, C., & Zhu, Y. (2021). Construction Industry Training Assessment Framework. Frontiers in Built Environment, 7(August), 1–12. https://doi.org/10.3389/fbuil.2021.678366

Jann, B., & Hinz, T. (2017). Research Question and Design for Survey Research In : The SAGE Handbook of Survey Methodology. 105–121.

Jih, H. J. (1996). The impact of learners' pathways on learning performance in multimedia Computer Aided Learning. Journal of Network and Computer Applications, 19(4), 367–380. https://doi.org/10.1006/jnca.1996.0025

Kelly, M., Lyng, C., McGrath, M., & Cannon, G. (2009). A multi-method study to determine the effectiveness of, and student attitudes to, online instructional videos for teaching clinical nursing skills. Nurse Education Today, 29(3), 292–300. https://doi.org/10.1016/j.nedt.2008.09.004

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2008). Evaluating Training Programs - The four levels (3rd editio). Berrett-Koehler Publishers, Inc.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2014). Implementing the Four Levels: a practical guide for effective evaluation of training programs. In Springer Series in Optical Sciences (Vol. 184).

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

Krötz, M., & Deutscher, V. (2021). Differences in Perception Matter – How Differences in the Perception of Training Quality of Trainees and Trainers Affect Drop-Out in VET. In Vocations and Learning (Vol. 14, Issue 3). https://doi.org/10.1007/s12186-021-09263-7

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The Validity of Self-Reported Grade Point Averages , Class Ranks , and Test Scores : A Meta- Analysis and Review of the Literature Author (s): Nathan R . Kuncel , Marcus Credé and Lisa L . Thomas Published by : American Educational Research Association. American Educational Research Association, 75(1), 63–82.

Lawlor, K. B. (2012). Smart Goals: How the Application of Smart Goals can Contribute to Achievement of Student Learning Outcomes. Journal of Developments in Business Simulation and Experiential Learning, 39, 259–267.

Markiewicz, A., & Patrick, I. (2016). Developing Monitoring and Evaluation Frameworks. SAGE Publications, Inc. https://doi.org/10.4135/9781071878774

Marzano, R. J. (2009). Designing & teaching learning goals & objectives. Marzano Research Laboratory.

Mathison, S. (2011). Encyclopedia of Evaluation. In Sage Research Methods. Sage Publications, Inc. https://doi.org/10.4135/9781412950558

Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution. Klagenfurt. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

McKenney, S., & Reeves, T. C. (2012). Conducting Educational Design Research. In Conducting Educational Design Research. https://doi.org/10.4324/9781315105642

McKenney, S., & Reeves, T. C. (2021). Educational design research: Portraying, conducting, and enhancing productive scholarship. Medical Education, 55(1), 82–92. https://doi.org/10.1111/medu.14280

Mcnall, M. A. (2018). SAGE Research Methods The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation Goals and Objectives. https://doi.org/10.4135/9781506326139.n289

Merchant, K. A., Stringer, C., & Theivananthampillai, P. (2010). Relationships between objective and subjective performance ratings. University of Southern California. Department of Accountancy and Business Law. Working paper. November 2009.

Newcomer P. W., K. E. & W. (2004). Handbook of Practical Program Evaluation. In Handbook of Practical Program Evaluation. https://ebookcentral.proquest.com/lib/unimelb/detail.action?docID=624388

Nielsen, K., & Shepherd, R. (2022). Understanding the outcomes of training to improve employee mental health: A novel framework for training transfer and effectiveness evaluation. Work and Stress, 1–15. https://doi.org/10.1080/02678373.2022.2028318

Oliver, D., Dobele, T., Greber, M., & Roberts, T. (2004). This Course has a Bloom Rating of 3.9. Proceedings of the Sixth Australasian Conference on Computing Education, 227–231.

Özbiçakçi, Ş., Bilik, Ö., & Intepeler, Ş. S. (2012). Assessment of goals in problem-based learning. Nurse Education Today, 32(8). https://doi.org/10.1016/j.nedt.2012.03.017

Paavola, S., & Hakkarainen, K. (2005). The knowledge creation metaphor - An emergent epistemological approach to learning. Science and Education, 14(6), 535–557. https://doi.org/10.1007/s11191-004-5157-0

Park, S., & Lee, J. Y. (2018). Workplace learning measures for human resource development: review and summary. Industrial and Commercial Training, 50(7–8), 420–431. https://doi.org/10.1108/ICT-08-2018-0068

Pike, G. R. (1995). The relationship between self-reports of college experiences and achievement test scores. Research in Higher Education, 36(1), 1–21. https://doi.org/10.1007/BF02207764

Pilli, O., & Admiraal, W. (2017). Students' Learning Outcomes in Massive Open Online Courses (MOOCs): Some Suggestions for Course Design. Yuksekogretim Dergisi, 7(1), 46–71. https://doi.org/10.2399/yod.17.001

Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. Educational Psychology Review, 16(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x

Pool, I. A., Poell, R. F., Berings, M. G. M. C., & ten Cate, O. (2016). Motives and activities for continuing professional development: An exploration of their relationships by integrating literature and interview data. Nurse Education Today, 38, 22–28. https://doi.org/10.1016/j.nedt.2016.01.004

Post, L. S., Guo, P., Saab, N., & Admiraal, W. (2019). Effects of remote labs on cognitive, behavioral, and affective learning outcomes in higher education. Computers and Education, 140(December 2018). https://doi.org/10.1016/j.compedu.2019.103596

Pusateri, T., Halonen, J. S., Hill, B., & McCarthy, M. (2009). The assessment cyberguide for learning goals and outcomes. November, 120. http://digitalcommons.kennesaw.edu/facpubs/433/

Reio, T., & Ghosh, R. (2009). Antecedents and Outcomes of Workplace Incivility. Computational Complexity, 2(1), 1–9. https://doi.org/10.1002/hrdq

Rogaten, J., & Rienties, B. (2021). A critical review of learning gains methods and approaches. International Perspectives on Higher Education Research, 14, 17–31. https://doi.org/10.1108/S1479-362820210000014003

Rolheiser, C., & Ross, J. A. (2001). Student self-evaluation: What research says and what practice shows. Plain Talk about Kids, 43, 57.

Ruiz-Calleja, A., Prieto, L. P., Ley, T., Rodríguez-Triana, M. J., & Dennerlein, S. (2021). Learning Analytics for Professional and Workplace Learning: A Literature Review. IEEE Transactions on Learning Technologies, 14(3), 353–366. https://doi.org/10.1109/TLT.2021.3092219

Russ-Eft, D. (2002). A Typology of Training Design and Work Environment Factors Affecting Workplace Learning and Transfer. Human Resource Development Review, 1(1), 45–65. https://doi.org/10.1177/1534484302011003

Saks, A. M., & Belcourt, M. (2006). An investigation of training activities and transfer of training in organisations. Human Resource Management, 45(4), 629–648. https://doi.org/10.1002/hrm.20135

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. Annual Review of Psychology, 52, 471–499. https://doi.org/10.1146/annurev.psych.52.1.471

Schiekirka, S., Reinhardt, D., Beibarth, T., Anders, S., Pukrop, T., & Raupach, T. (2013). Estimating learning outcomes from pre-and posttest student self-assessments: A longitudinal study. Academic Medicine, 88(3), 369–375. https://doi.org/10.1097/ACM.0b013e318280a6f6

Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the use of self-generated identification codes. Evaluation Review, 34(5), 391–418. https://doi.org/10.1177/0193841X10387576

Schunk, D. H. (2010). Reading &Writing Quarterly SELF-EFFICACY FOR READING AND WRITING: INFLUENCE OF MODELING, GOAL SETTING, AND SELF-EVALUATION. https://doi.org/10.1080/10573560308219

Scriven, M. (1994). Evaluation as a discipline. Studies in Educational Evaluation, 20(1), 147–166. https://doi.org/10.1016/S0191-491X(00)80010-3

Sedikides, C., & Strube, M. J. (1997). Self-Evaluation: To Thine Own Self Be Good, To Thine Own Self Be Sure, To Thine Own Self Be True, and To Thine Own Self be Better. Advances in Experimental Social Psychology, 29, 209–269. https://doi.org/00672601197

Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. Educational Researcher, 27(2), 4. https://doi.org/10.2307/1176193

Silva, M., & Paulo Rodrigues, P. (2020). DIDO requirements analysis report. https://drive.google.com/file/d/11KUQS3p9iaEv84hzPYxJ9FPnR7NgxfO4/view

Smidt, A., Balandin, S., Sigafoos, J., & Reed, V. A. (2009). The Kirkpatrick model: A useful tool for evaluating training outcomes. Journal of Intellectual and Developmental Disability, 34(3), 266–274. https://doi.org/10.1080/13668250903093125

So, H. J., & Brush, T. A. (2008). Student perceptions of collaborative learning, social presence, and satisfaction in a blended learning environment: Relationships and critical factors. Computers and Education, 51(1), 318–336. https://doi.org/10.1016/j.compedu.2007.05.009

Steiner, G. (2001). Transfer of Learning, Cognitive Psychology of. In International Encyclopedia of the Social & Behavioral Sciences (pp. 15845–15851). Pergamon. https://doi.org/10.1016/b0-08-043076-7/01481-9

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organisations. Annual Review of Psychology, 43(1), 399–441. https://doi.org/10.1146/ANNUREV.PS.43.020192.002151

Thorndyke, L. E., Gusic, M. E., & Milner, R. J. (2008). Functional mentoring: A practical approach with multilevel outcomes. Journal of Continuing Education in the Health Professions, 28(3), 157–164. https://doi.org/10.1002/chp.178

Tynjälä, P. (2013). Toward a 3-P Model of Workplace Learning: A Literature Review. Vocations and Learning, 6(1), 11–36. https://doi.org/10.1007/s12186-012-9091-z

Tzeng, G. H., Chiang, C. H., & Li, C. W. (2007). Evaluating intertwined effects in e-learning programs: A novel hybrid MCDM model based on factor analysis and DEMATEL. Expert Systems with Applications, 32(4), 1028–1044. https://doi.org/10.1016/j.eswa.2006.02.004

van der Schaaf, M., Donkers, J., Slof, B., Moonen-van Loon, J., van Tartwijk, J., Driessen, E., Badii, A., Serban, O., & Ten Cate, O. (2017). Improving workplace-based assessment and feedback by an E-portfolio enhanced with learning analytics. Educational Technology Research and Development, 65(2), 359–380. https://doi.org/10.1007/s11423-016-9496-8

van Uum, M. S. J., & Pepin, B. (2022). Students' self-reported learning gains in higher engineering education. European Journal of Engineering Education, 1–17. https://doi.org/10.1080/03043797.2022.2046708

Venkatesh, Morris, Davis, & Davis. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425. https://doi.org/10.2307/30036540

Wang, Y. S. (2003). Assessment of learner satisfaction with asynchronous electronic learning systems. Information and Management, 41(1), 75–86. https://doi.org/10.1016/S0378-7206(03)00028-4

Wanzer, D. L. (2021). What Is Evaluation?: Perspectives of How Evaluation Differs (or Not) From Research. American Journal of Evaluation, 42(1), 28–46. https://doi.org/10.1177/1098214020920710

Wei, X., Saab, N., & Admiraal, W. (2021). Assessment of cognitive, behavioral, and affective learning outcomes in massive open online courses: A systematic literature review. Computers and Education, 163(October 2020), 104097. https://doi.org/10.1016/j.compedu.2020.104097

Wiese, A., Kilty, C., Bergin, C., Flood, P., Fu, N., Horgan, M., Higgins, A., Maher, B., O'Kane, G., Prihodova, L., Slattery, D., & Bennett, D. (2017). Protocol for a realist review of workplace learning in postgraduate medical education and training. Systematic Reviews, 6(1), 1–6. https://doi.org/10.1186/s13643-017-0415-9

Yan, Z. (2019). Assessment & Evaluation in Higher Education Self-assessment in the process of self-regulated learning and its relationship with academic achievement. https://doi.org/10.1080/02602938.2019.1629390

8 Appendices

APPENDIX A

Coding Scheme for Document Analysis

Main Theme	Subject Item	Definition
Hinderances	1. Lack of Awareness	1. Many stakeholders may be unaware of the existence or benefits of the rubric.
	2. Resistance to Change	2. Stakeholders may resist adopting the rubric due to concerns about disruption, unfamiliarity, or perceived risks associated with changing existing practices.
	3. Complexity and Difficulty	3. The rubric may be perceived as complex or challenging to understand, implement, or integrate into existing processes.
	4. Resource Constraints	4. Limited availability of financial, human, time, or technological resources.
	5. Inadequate Training and Support	5. Insufficient training and support for stakeholders in understanding and using the rubric.
	6. Compatibility Issues	6. The rubric may not align well with the existing organisational culture, processes, or systems.
	7. Lack of Stakeholder Engagement	7. Insufficient involvement and engagement of key stakeholders in the development and implementation of the rubric.
	8. Perceived Irrelevance or Ineffectiveness	8. Stakeholders may perceive the rubric as irrelevant to their specific context or industry, or they may doubt its effectiveness in achieving desired outcomes.
	9. Legal and Regulatory Barriers	9. Existing laws, regulations, or policies may pose barriers to the adoption of the rubric.
	10. Other	10.
Main Theme	**Subject Item**	**Definition**
Capabilities	11. Expertise in Evaluation Design	11. Strong understanding of evaluation principles, methodologies, and best practices.
	12. Skilled Evaluation Team	12. Experienced evaluators with the necessary knowledge and skills to design effective evaluation programs.
	13. Resources	13. Access to the financial, human, time, and technological resources required to support evaluation program design.
	14. Data Collection and Analysis	14. Ability to collect and analyse data
	15. Adaptability and Flexibility	15. Ability to adapt evaluation designs to fit specific contexts and respond to evolving needs and circumstances.
	16. Planning and Implementation	16. Skills to develop comprehensive evaluation plans and effectively implement evaluation activities.
	17. Other	17.
Main Theme	**Subject Item**	**Definition**

Limitations	18. Expertise in Evaluation Design	18. Strong understanding of evaluation principles, methodologies, and best practices.
	19. Skilled Evaluation Team	19. Experienced evaluators with the necessary knowledge and skills to design effective evaluation programs.
	20. Resources	20. Access to the financial, human, time, and technological resources required to support evaluation program design.
	21. Data Collection and Analysis	21. Ability to collect and analyse data
	22. Adaptability and Flexibility	22. Ability to adapt evaluation designs to fit specific contexts and respond to evolving needs and circumstances.
	23. Planning and Implementation	23. Skills to develop comprehensive evaluation plans and effectively implement evaluation activities.
	24. Other	24.
Main Theme	**Subject Item**	**Definition**
Needs	25. Training and Capacity Building	25. Require training or resources to develop the necessary knowledge, skills, and competencies in evaluation design.
	26. Access to Expertise	26. Require access to experts or consultants who can provide guidance and support in designing evaluation programs.
	27. Clear Guidelines and Frameworks	27. Require clear guidelines, frameworks, or best practices that provide a structured approach to designing evaluation programs.
	28. Collaboration and Knowledge Sharing	28. Require opportunities to collaborate with peers, share experiences, and learn from each other's evaluation program design processes.
	29. Adequate Resources	29. Require sufficient financial, human, time, and technological resources to design and implement evaluation programs.
	30. Data Collection and Analysis	30. Require assistance in data collection, management, and analysis to support the design and implementation of evaluation programs.
	31. Evaluation Tools and Methodologies	31. Require access to evaluation tools, methodologies, and instruments that are suitable for their specific context and objectives.
	32. Flexibility and Adaptability	32. Require the ability to adapt and customise evaluation program designs to fit their unique needs and circumstances.
	33. Other	33.

APPENDIX B

Outcomes of Interview Research

Summary of all countries

	Netherlands	Bulgaria	Spain	Austria	France	Ireland
Upskilling activities	*Supply side* Cross-trade EE skills, coordination of work among different craftworkers (?) BUS app E-learning, micro learning modules (via BUS app with QR) Experimentation in training centres; monthly, f2f *Demand side* DIY stores (work in progress, not usable)	*Supply side* Hubs: specialised training centres for designers and construction workers (trained trainers), no data is recorded. Online NZEB course for specialists and professionals. Large platform with online trainings (can be combined with f2f training).	*Supply side* Installers (that collaborate with Bauhaus) can also participate, but comes down to pro-active behaviour and motivated attitude. *Demand side* E-pills: short electronic messages for Bauhaus staff, to be able to sell the products. Participation from staff is mandatory.	*Supply side* App. *Demand side* Short trainings with energy experts that go by family houses as consultants (ARGE Eba is main stakeholder).	*Supply side* AFEST: Continuous education BPT: The experimental project “Bâtis Ton Projet – BTP” (Build your project): raise awareness among the public about the different trades of the construction sector, but also to evaluate and identify potential candidates	*Supply side* Impromptu (on site, ad hoc) for hardware stores that focus on how to get contractors on board, how to do e.g. airtightness and what materials to use Small online training, micro trainings (see Erasmus+ construction blueprint) *Demand side* Staff training: step-by-step training and short awareness campaigns so that staff can advice home owners

Questions, needs, comments	*Supply side* For participants: To what extent is learning in trainings/courses/workshops transferred to workers' daily practice. E.g., how can new skills be applied on construction site. The value of informal learning. E.g., does e-learning lead to up-to-date knowledge and does it motivate people. *Demand side* For customers: How can we show trainings are effective? This helps procurement. Effective characteristics of upskilling activities can be made a requirement by grant provider (e.g., RVO). *Observed challenges* Do upskilling activities boost craftworkers' appreciation of upskilling? Motivation of craftworkers is generally low (they don't think they need it). How to stimulate participation. For example, address the BUS-app in value appraisals by managers. *Design characteristics* Work schedule of 'bingo sheet' with design principles for upskilling interventions and tips on how to evaluate upskilling interventions (e.g., experimental of control group or make photos).	*Observed challenges* Lack of CPD systems for construction workers. Aim is to design this (different learning outcomes are already systemised in WP2). Satisfaction surveys are not conducted because participants presence and remarks are enough.	*Observed challenges* Bauhaus probably check effectiveness themselves (e.g., are products sold more? Through satisfaction surveys?). *Design characteristics* Consider the profile of blue-collar workers: it should be engaging, attractive, short, fast, direct.	*Observed challenges* Focus on stimulating demand side (consultants) because construction companies do not see the added value of upskilling (is too expensive, takes too much time). Blue collar workers will not use app outside work. Blue collar workers prefer to learn by doing on the side, not online or in app (this is certainly the case for 'inleners' who have had different education White collar workers are more willing to use apps, because of experience and genuine interest. *Design characteristics* Upskilling sapp hould be userfriendly, learning gains should be visible to users and their employers, up to date so that it stimulates CPD.	*Observed challenges* Target population is craftworkers and administrators Mostly f2f Motivation is generally low, experienced craftworkers believe that they are done learning	*Observed challenges* Hardware stores/association should be involved to encourage contractors to do certified trainings People do not take the training. More need for blended, self-directing etc. Target group is staff of hardware stores, local authorities, and construction workers *Design characteristics* Consider usability, WIFI access, not using too large files, demand, and awareness, simple language.

From: *Evaluation framework of upskilling in the construction sector V0.5 (2021).*

APPENDIX C

Briefing and Debriefing Forms

UNDERSTANDING THE TYPE OF TRAINING

Institution or Organisation	**Your name**

Training	
Focus of training	
Target audience:	*Number of Participants:*

Provider of training	
☐ Own institution or organisation ☐ Commissioned training provider	☐ Partner institution or organisation ☐ Other:

Type of Training		☐ Voluntary ☐ Mandatory
☐ In-class ☐ On-site ☐ Online	☐ Hybrid ☐ Other:	**Duration and length of training** (e.g., *one-hour online training, full-day workshop, part of a modular program, etc.)*

Elements contained in the training program (multiple items may apply)			
Offline		**Online**	
☐ face-to-face instruction ☐ reading material ☐ hands-on practice	☐ Q&A session ☐ demonstration ☐ other:	☐ video lessons ☐ e-learning ☐ discussion forum	☐ online tutorial ☐ Other:

Additional information/ comments

Design principles (Saks & Belcourt, 2006)	+	-
Pretraining activities		
1. Trainee input and involvement		
1.1. Employees are given advanced notification about training content prior to attending a training program.		
1.2. Training needs of employees are identified prior to training.		
1.3. Employees have pre-course discussions with their supervisors prior to attending a training program.		
1.4. Employees have input in decisions about training program content and/or methods.		
2. Supervisor involvement		
2.1. Employees are given release time to prepare for a training program.		
2.2. Supervisors discuss the content and benefits of a training program with employees prior to a training program.		
2.3. Supervisors participate in advance orientation or training sessions regarding the training programs to which they will send their employees.		
2.4. Supervisors set goals with employees that focus on improving specific skills before employees attend training programs.		
3. Training attendance policy		
3.1. Employees have a choice as to whether or not they will attend any particular training program.		
3.2. Attendance at training programs is voluntarily		
3.3. Employees from the same department or functional group are trained together.		
4. Trainee preparation		
4.1. Trainees are given preparatory reading prior to attending a training program		
4.2. Training programs include activities or assignments that trainees are required to do before they arrive for the actual training program.		
During training activities		
5. Support		
5.1. Training programs provide trainees with training experiences and conditions (surroundings, tasks, equipment) that closely resemble those in the actual work environment.		
5.2. Training programs provide trainees with a variety of training stimuli and experiences, such as several examples of a concept, or practice experiences in a variety of situations.		
5.3. Training programs teach trainees the general rules and theoretical principles that underlie the training content and the use and application of the trained skills.		
5.4. Trainees are given feedback and information about their performance of the training tasks and material during the training program.		
5.5. Trainees are rewarded during training for learning and performing training material and tasks.		
5.6. Trainees leave training programs with a written performance contract with goals to be achieved.		
5.7. Training programs prepare trainees to cope with obstacles or difficulties that might prevent them from successfully applying the training material when they return to the work environment.		
Post training activities		
6. Supervisor support		

6.1. Supervisors are instructed to provide trainees with support to help them use newly acquired skills after attending a training program.		
6.2. Supervisors are instructed to ensure that trainees have opportunities to practice and apply newly acquired knowledge and skills after attending a training program.		
6.3. Supervisors are instructed to praise or reward employees for using newly acquired skills developed in a training program.		
7. Organisation support		
7.1. Some form of booster session is conducted as an extension of a training program in which the trainer meets with trainees.		
7.2. Efforts are made to ensure that employees have the resources (e.g., tools, equipment, materials, supplies, etc.) that are necessary in order to apply the knowledge, skills, and/or abilities developed in training programs.		
7.3. The performance appraisal system considers trainees' use of knowledge, skills, and/or abilities acquired in training programs.		
8. Accountability		
8.1. Trainees are required to submit a post-training report after attending a training program.		
8.2. Trainees are required to participate in an interview or discussion as part of a follow-up to a training program they attended.		
9. Evaluation and feedback		
9.1. Employees are paired with each other following completion of a training program in order to assist each other by providing feedback and reinforcement to ensure they use the skills developed in a training program.		
9.2. Employees are evaluated on their use of new skills or knowledge following completion of a training program.		
9.3. Employees are required to undergo an assessment following completion of a training program in order to evaluate their learning.		

LEARNING GOALS

	WHEN	WHO		Level of Learning:	WHAT
1.			***will be able to***		
2.			***will be able to***		
3.			***will be able to***		
4.			***will be able to***		

DEBRIEFING

Initial Expectations
1. What were the expectations related to the training?
2. What were the expectations related to the evaluation in terms of:
2.1. Results
2.2. Information collected (data)
3. What type of information are you expecting to obtain?
4. How would you normally obtain these data?
5. What is the usefulness of such data for decision-making processes?
6. What kind of decisions are influenced by such data?
7. Is this training aiming for a certification?
Results
8. What did you learn from the data obtained?
9. Do this type of information influence decisions for future interventions?
10. Was all the expected information collected?
10.1. What information was not collected?
11. Which information is the most relevant?
11.1. And the least relevant?
12. What other information do you think would be worth collecting and analysing?
Evaluation Process
13. What did you think of the evaluation process?
13.1. Timing
13.2. Costs
13.3. Effort (Human resources)
14. Did you experience any difficulty in the process?
14.1. designing an evaluation
14.2. implementing the evaluation
14.3. collecting the results
15. What is your perception on its usefulness?
15.1. Pros:
15.2. Cons:
16. Is it feasible to adopt this type of assessment in a realistic context?
17. And do you think it could also be suitable within the BUSLeague context?
Support
General overview of tools
18. What were your expectations in terms of support for developing an evaluation?
19. How do you think the tools provided helped you implement the evaluation?
20. Where do you think support was insufficient?
21. If you were to implement this by yourself in the future, what kind of support would be needed?
22. Support for the planning of an evaluation (time, participants, methodology)

22.1.	Why?
22.2.	How would you want this support to be?
23. Support for the structuring of an evaluation (L1,L2 and L3)	
23.1.	Why?
23.2.	How would you want this support to be?
24. Support for selection of tools (pre-posttests, different instruments)	
24.1.	Why?
24.2.	How would you want this support to be?
25. Support for data analysis	
25.1.	Why?
25.2.	How would you want this support to be?
26. Other:	
26.1.	Why?
26.2.	How would you want this support to be?
Learning Goals Tool	
27. Do you think the layout of the tool is clear?	
28. Do you think the tool is easy to use/user friendly?	
29. Do you think the instructions contained in the Learning Goals tool are sufficient?	
30. What difficulties have you had in using the tool?	
31. Do you think the tool is useful for its purposes?	
32. How do you think the tool could be improved?	
Reference Guide	
33. Do you think the layout of the reference guide is clear?	
34. Do you think the reference guide is easy to use/user friendly?	
35. Do you think the text in the reference guide is easy to understand?	
36. Do you think the instructions contained in the reference guide are sufficient?	
37. What difficulties have you had in using the reference guide?	
37.1.	Structuring an evaluation
37.2.	Evaluation123 - decision tree
37.3.	Level 1 - quality assessment
37.4.	Level 1- design principles
37.5.	Educational effectiveness
37.6.	Level 1 – reaction
37.7.	Level 2 - learning outcomes
37.8.	Designing knowledge and skills tests
37.9.	Self-assessment of learning goals
37.10.	Determining learning goals
37.11.	Using the learning goal tool
37.12.	Operationalisation of learning goals assessments
37.13.	Level 3 – transfer of learning
37.14.	Exploring paths and data
37.15.	Learning paths
37.16.	Learning analytics
37.17.	Energy performance gap (EPG)

38. Do you think the reference guide is useful for its purposes?
39. How do you think the reference guide could be improved?
Question Packs
40. Do you think the organisation of the Question packs is clear?
41. Do you think items in the Question Packs are easy to understand?
42. Do you think the Question Packs are easy to use/user friendly?
43. What difficulties have you had in using the Question Packs?
44. Do you think the Question Packs are useful for its purposes?
45. How do you think the Question Packs could be improved?
Decision Tree
46. Do you think the layout of the decision tree is clear?
47. Do you think the decision tree is easy to use/user friendly?
48. Do you think the instructions contained in the decision tree tool are sufficient?
49. What difficulties have you had in using the decision tree?
50. Do you think the decision tree is useful for its purposes?
51. How do you think the decision tree could be improved?
52. **Open Question**

Design solution 1: Creation of Reference Guide

Inputs	Activities	Outputs	Outcomes	Impact
• Existing project documentation and evaluation rubric • Expertise in evaluation design and construction industry knowledge	• Identify and gather relevant information, instructions, and guidance on evaluation design and construction industry-specific considerations. • Condense and organize the collected knowledge into a comprehensive and user-friendly format. • Develop a manual that offers clear explanations, step-by-step instructions, and troubleshooting tips for designing structured evaluations. • Review and validate the manual's content through expert input and feedback.	• Comprehensive and organised manual on evaluation design for the construction industry. • User-friendly resource tool that provides valuable information, instructions, and guidance.	• Increased accessibility to knowledge and expertise on evaluation design for evaluators in the construction industry. • Improved understanding and utilisation of the evaluation rubric. • Enhanced ability to design structured evaluations in a systematic and effective manner. • Increased confidence, efficiency, and optimal performance of evaluators.	• Increased accessibility to knowledge and expertise on evaluation design for evaluators in the construction industry. • Improved understanding and utilisation of the evaluation rubric. • Enhanced ability to design structured evaluations in a systematic and effective manner. • Increased confidence, efficiency, and optimal performance of evaluators.

Design solution 2: Creation of a Tool for Structured Formulation of Learning Goals.

Inputs	Activities	Outputs	Outcomes	Impact
• Knowledge and expertise on learning goal formulation • SMART criteria for clear and measurable goals • Bloom's taxonomy for appropriate verb selection • Piloting observations and feedback from evaluators	• Research and gather information on effective learning goal formulation strategies. • Develop a guide/tool that integrates the SMART criteria and Bloom's taxonomy for structured formulation of learning goals. • Design the guide/tool in a digital format, such as a spreadsheet, to facilitate ease of use and minimize errors. • Test and refine the guide/tool based on feedback and pilot testing.	• Guide/tool for structured formulation of learning goals, incorporating SMART criteria and Bloom's taxonomy. • Digital format (e.g., spreadsheet) that provides an organised and structured framework for goal formulation.	• Improved ability of evaluators to formulate clear, measurable, and meaningful learning goals. • Enhanced understanding and application of the SMART criteria and Bloom's taxonomy in goal formulation. • Reduction in incomplete or unmeasurable learning goal statements. • Increased efficiency and accuracy in the operationalisation of learning goals.	• Improved quality and effectiveness of educational programs and interventions. • Enhanced alignment between learning goals and desired outcomes. • Facilitation of assessment and evaluation processes by providing well-defined and measurable learning goals. • Increased confidence and competence among evaluators in formulating effective learning goals.

Design solution 3: Structuring of Survey Items

Inputs	Activities	Outputs	Outcomes	Impact
• Valid survey instruments related to the first three levels of evaluation (Rubric) • Feedback from the Analysis and Exploration phase • Technological affordances for survey organisation, data collection, and analysis	• Examine and categorize the valid survey instruments based on their specific aspects and constructs (e.g., satisfaction with training, quality of teaching, normative success). • Develop explanations and instructional materials to educate evaluators about the utility and operationalisation of survey instruments. • Create guidelines for prioritizing the collection of data that can be easily visualised and descriptively analysed. • Incorporate technological affordances to facilitate survey organisation, data collection, and treatment processes.	• Categorised survey instruments based on specific evaluation aspects and constructs. • Explanations and instructional materials for evaluators on the utility and operationalisation of survey instruments. • Guidelines for prioritizing the collection of data that allows for easy visualisation and descriptive analysis. • Technological tools or platforms to support survey organisation, data collection, and treatment.	• Increased knowledge and understanding among evaluators regarding the purpose and application of different survey instruments. • Improved ability to select and prioritize survey instruments based on evaluation needs and desired outcomes. • Enhanced skills in the collection and treatment of survey data, including measuring Likert scale items and compiling data. • Efficient use of technology to streamline survey processes, data collection, and analysis.	• Enhanced quality and reliability of evaluation data collected through surveys. • Improved ability to visualize and analyse survey data in a descriptive manner. • Enhanced capacity for evidence-based decision-making and program improvement. • Increased confidence and competence among evaluators in conducting evaluations and utilizing survey instruments effectively.

APPENDIX E

Focus Group Guided Interview

Evaluation123
1. What did you think of the evaluation process?
1.1. Timing
1.2. Costs
1.3. Effort (Human resources)
2. Did you experience any difficulty in the process?
2.1. designing an evaluation
2.2. implementing the evaluation
2.3. collecting the results
3. What is your perception on its usefulness?
3.1. Pros:
3.2. Cons:
4. Is it feasible to adopt this type of assessment in a realistic context?
5. And do you think it could also be suitable within the BUSLeague context?
Learning Goals Tool
6. Do you think the layout of the tool is clear?
7. Do you think the tool is easy to use/user friendly?
8. Do you think the instructions contained in the Learning Goals tool are sufficient?
9. What difficulties have you had in using the tool?
10. Do you think the tool is useful for its purposes?
11. How do you think the tool could be improved?
Reference Guide
12. Do you think the layout of the reference guide is clear?
13. Do you think the reference guide is easy to use/user friendly?
14. Do you think the text in the reference guide is easy to understand?
15. Do you think the instructions contained in the reference guide are sufficient?

16. What difficulties have you had in using the reference guide?
16.1. Structuring an evaluation
16.2. Evaluation123 - decision tree
16.3. Level 1 - quality assessment
16.4. Level 1- design principles
16.5. Educational effectiveness
16.6. Level 1 – reaction
16.7. Level 2 - learning outcomes
16.8. Designing knowledge and skills tests
16.9. Self-assessment of learning goals
16.10. Determining learning goals
16.11. Using the learning goal tool
16.12. Operationalisation of learning goals assessments
16.13. Level 3 – transfer of learning
16.14. Exploring paths and data
16.15. Learning paths
16.16. Learning analytics
16.17. Energy performance gap (EPG)
17. Do you think the reference guide is useful for its purposes?
18. How do you think the reference guide could be improved?
Question Packs
19. Do you think the organisation of the Question packs is clear?
20. Do you think items in the Question Packs are easy to understand?
21. Do you think the Question Packs are easy to use/user friendly?
22. What difficulties have you had in using the Question Packs?
23. Do you think the Question Packs are useful for its purposes?
24. How do you think the Question Packs could be improved?
Decision Tree
25. Do you think the layout of the decision tree is clear?

26. Do you think the decision tree is easy to use/user friendly?
27. Do you think the instructions contained in the decision tree tool are sufficient?
28. What difficulties have you had in using the decision tree?
29. Do you think the decision tree is useful for its purposes?
30. How do you think the decision tree could be improved?
General overview of tools
31. What were your expectations in terms of support for developing an evaluation?
32. How do you think the tools provided helped you implement the evaluation?
33. Where do you think support was insufficient?
34. If you were to implement this by yourself in the future, what kind of support would be needed?
34.1. Support for the planning of an evaluation (time, participants, methodology)
34.1.1. Why?
34.1.2. How would you want this support to be?
34.2. Support for the structuring of an evaluation (L1,L2 and L3)
34.2.1. Why?
34.2.2. How would you want this support to be?
34.3. Support for selection of tools (pre-post-tests, different instruments)
34.3.1. Why?
34.3.2. How would you want this support to be?
34.4. Support for data analysis
34.4.1. Why?
34.4.2. How would you want this support to be?
34.5. Other:
34.5.1. Why?
34.5.2. How would you want this support to be?
35. **Open Question**

UNIVERSITY OF TWENTE.

This survey aims at assessing the quality and improving the EVALUATION123 Reference Guide. This is an anonymous survey, and this data will be used and treated under the approval of the BMS Ethics Committee of the University of Twente - RN 220198

EVALUATION123

Please indicate to what extent you agree with the following statements regarding the EVALUATION123 Reference Guide

1. It is easy to look up things in the Reference Guide.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
2. The text in the Reference Guide is easy to read.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
3. The Reference Guide uses too many technical words.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
4. The images and layout in the Reference Guide facilitate comprehension of the topics.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
5. The Reference Guide provides a good overview of different possibilities for evaluations.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
6. The Reference Guide helps me understand my evaluation needs.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
7. The Reference Guide helps me understand my evaluation possibilities.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
8. I feel able to plan an evaluation for a training program using the Reference Guide.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
9. The Reference Guide will be useful for planning evaluations in different contexts.
   Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
10. The Decision Tree in the Reference Guide facilitates planning an evaluation.
    Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
11. The Decision Tree in the Reference Guide helps me aligning my efforts to my intentions for evaluations, getting the maximal evaluation outcomes within my possibilities.
    Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree
12. The Learning Goals Tool helps me specifying the ULOs (*units of learning outcomes*) to units that can be evaluated.
    Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree

13. The Learning Goals Tool helps me clarifying learning objectives.

Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree

14. The Learning Goals Tool helps me designing my evaluation.

Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree

15. The Questionnaire Packs help me selecting items for my surveys.

Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree

16. Having the Questionnaire Packs available in digital format (e.g. *GForms*) facilitates their use.

Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree

17. The checklist used during the workshop was useful for planning an evaluation.

Totally Disagree ○ ○ ○ ○ ○ ○ Totally agree

Please answer these questions about the EVALUATION123

18. In the Evaluation123, the levels are independent to each other

☐ True ☐ False ☐ I do not know

19. Learning goals are central to the design of evaluations aiming certifications

☐ True ☐ False ☐ I do not know

20. Pretests are essential conditions for evaluations

☐ True ☐ False ☐ I do not know

21. Knowledge and skill tests are the only valid ways of assessing learning outcomes.

☐ True ☐ False ☐ I do not know

22. Solid evaluations always contain subjective and objective assessments.

☐ True ☐ False ☐ I do not know

23. Would you like to make any further comment / feedback / suggestion related to the Reference Guide? All inputs are really appreciated.

APPENDIX G

REPORT SURVEY WS3 CM5

A survey was carried out at the end of the workshop in the fifth Consortium Meeting. The aim of the survey was to collect quantitative data on the usability and feasibility of the Reference Guide and associated tools as guiding instruments for designing evaluation plans.

23 questions

- 17 aimed at assessing the RG and associated tools
  - 6 points Likert scale (totally disagree – totally agree)
- 4 knowledge-test questions
  - Based on de Jong suggestion. This tested whether people how much the material was effective in teaching content to people
- 1 open-ended question eliciting feedbacks/comments/suggestions

Question: given they did not have the chance to thoroughly read the material and not all of them had previous experience with the E123, can the results of the knowledge test be trusted?

The following disclaimer was used at the top of the survey sheet:

- This survey aims at assessing the quality and improving the EVALUATION123 Reference Guide. This is an anonymous survey, and this data will be used and treated under the approval of the BMS Ethics Committee of the University of Twente - RN 220198
- ***Please indicate to what extent you agree with the following statements regarding the* EVALUATION123 Reference Guide**

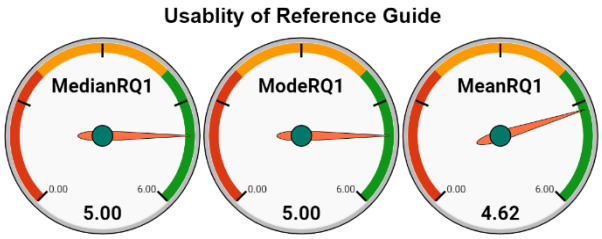
The 17 questions on the Reference Guide and associated tools were clustered in 6 themes, to clarify 6 (sub) research questions.

Key Statistics

	Valid N	Mean	Median	Mode	Std. Deviation
Usability RG	11	4,62	4,40	4,40	0,540
Usefulness RG	11	5,00	5,00	5,00	0,798
Usefulness DT	11	5,18	5,00	6,00	0,751
Usefulness LGT	11	4,52	4,67	4,00	0,848
Usefulness QP	11	5,14	5,00	5,50	0,393
Usefulness Checklist	11	4,91	5,00	5,00	0,701

Statistics	Mean	Med	Mode	S.D.	Min	Max
Q1 It is easy to look up things in the Reference Guide.	5.00	5	4-6[a]	.894	4	6
Q2 The text in the Reference Guide is easy to read.	4.36	4	4-5[a]	.674	3	5
Q3Rev The Reference Guide uses too many technical words	3.73	4	3	1.009	2	5
Q4 The images and layout in the Reference Guide facilitate comprehension of the topics.	4.91	5	5	.944	3	6
Q5 The Reference Guide provides a good overview of different possibilities for evaluations.	5.09	5	5	.944	3	6
Q6 The Reference Guide helps me understand my evaluation needs.	5.00	5	5	.894	3	6
Q7 The Reference Guide helps me understand my evaluation possibilities.	4.82	5	4-6[a]	1.079	3	6
Q8 I feel able to plan an evaluation for a training program using the Reference Guide.	4.91	5	5	.944	3	6
Q9 The Reference Guide will be useful for planning evaluations in different contexts.	5.27	5	6	.786	4	6
Q10 The Decision Tree in the Reference Guide facilitates planning an evaluation.	5.36	6	6	.809	4	6
Q11 The Decision Tree in the Reference Guide helps me aligning my efforts to my intentions for evaluations, getting the maximal evaluation outcomes within my possibilities.	4.90	5	4	.876	4	6
Q12 The Learning Goals Tool helps me specifying the ULOs (units of learning outcomes) to units that can be evaluated.	4.09	4	5	1.375	1	6
Q13 The Learning Goals Tool helps me clarifying learning objectives.	4.45	4	4	.820	3	6
Q14 The Learning Goals Tool helps me designing my evaluation.	5.00	5	5	.632	4	6
Q15 The Questionnaire Packs help me selecting items for my surveys.	4.73	5	5	.786	3	6
Q16 Having the Questionnaire Packs available in digital format (e.g., GForms) facilitates their use.	5.55	6	6	.522	5	6
Q17 The checklist used during the workshop was useful for planning an evaluation.	4.91	5	5	.701	3	6

1. Usability of the Reference Guide (Q1 – Q5)
RQ_1: The Reference Guide is easy to use.

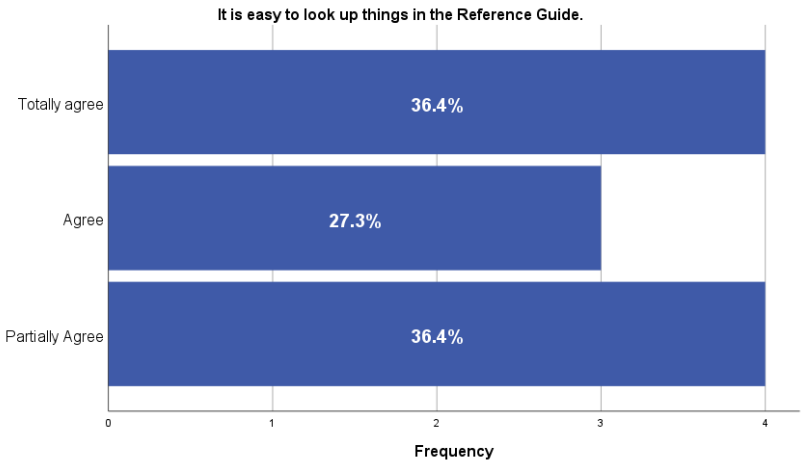


medianRQ1 Usability of Reference Guide

	N	%
4.00 Partially Agree	4	36.4%
5.00 Agree	5	45.5%
6.00 Totally agree	2	18.2%

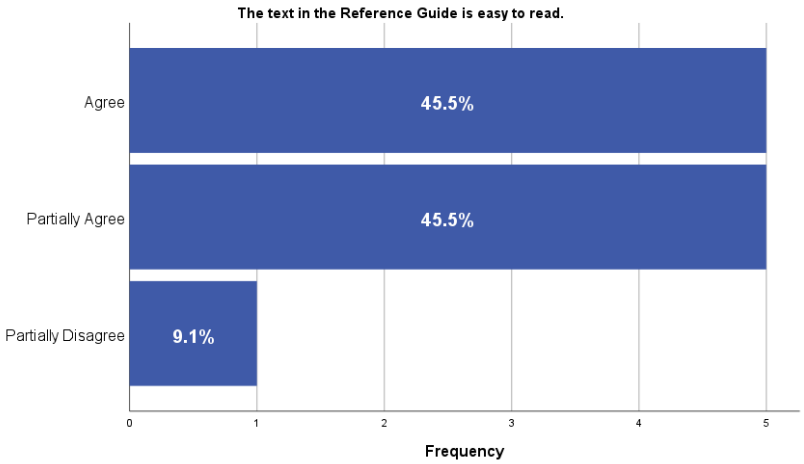
1.1. It is easy to look up things in the Reference Guide.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



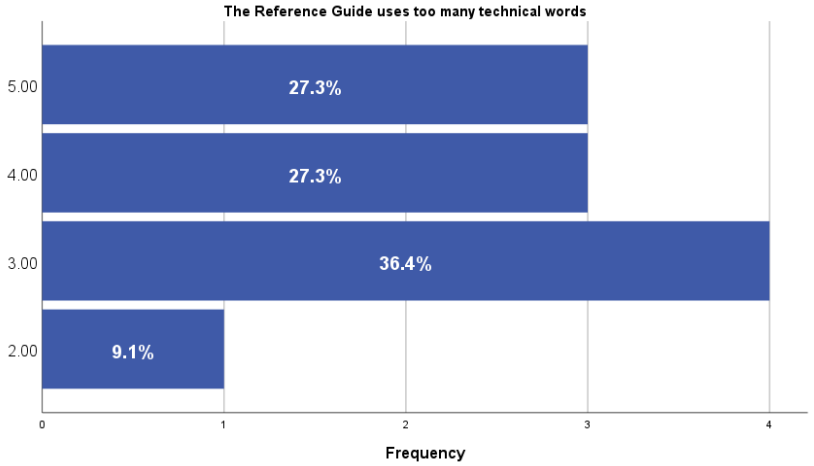
1.2. The text in the Reference Guide is easy to read.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



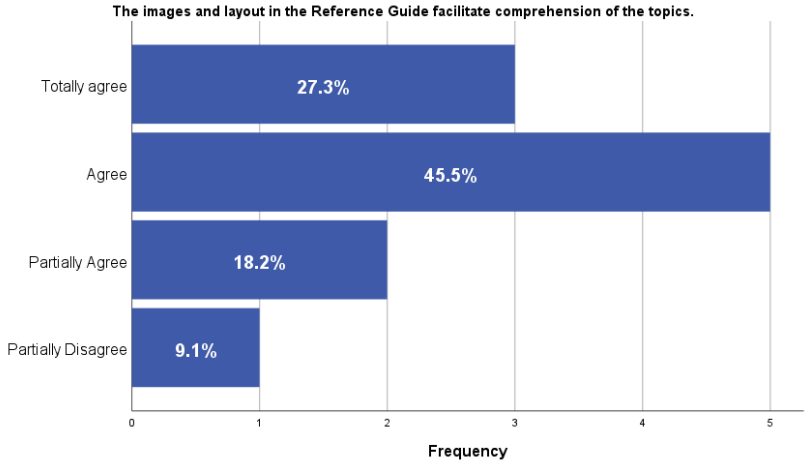
1.3. The Reference Guide uses too many technical words.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



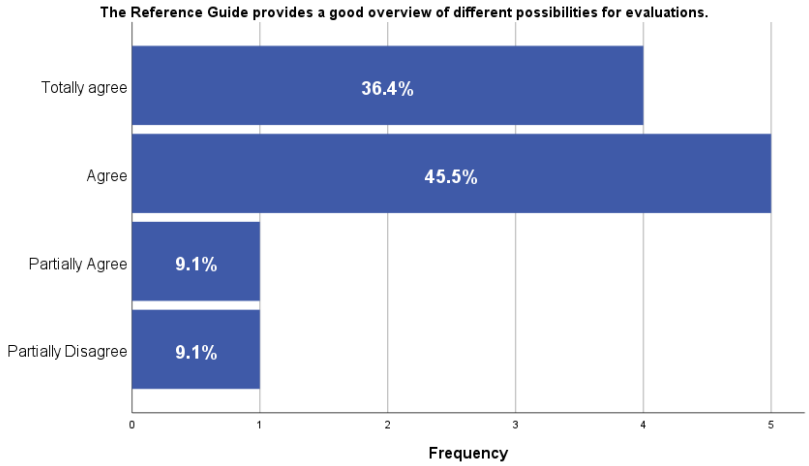
1.4. The images and layout in the Reference Guide facilitate comprehension of the topics.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



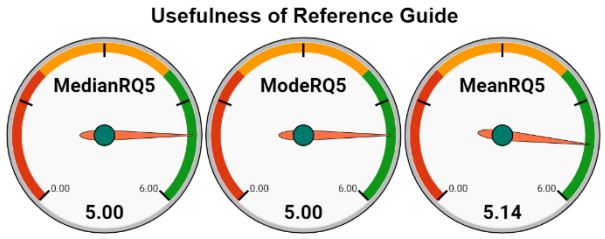
1.5. The Reference Guide provides a good overview of different possibilities for evaluations.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



2. Usefulness of the Reference Guide (Q6 – Q9)

RQ 2: The Reference Guide facilitates designing evaluation plans.

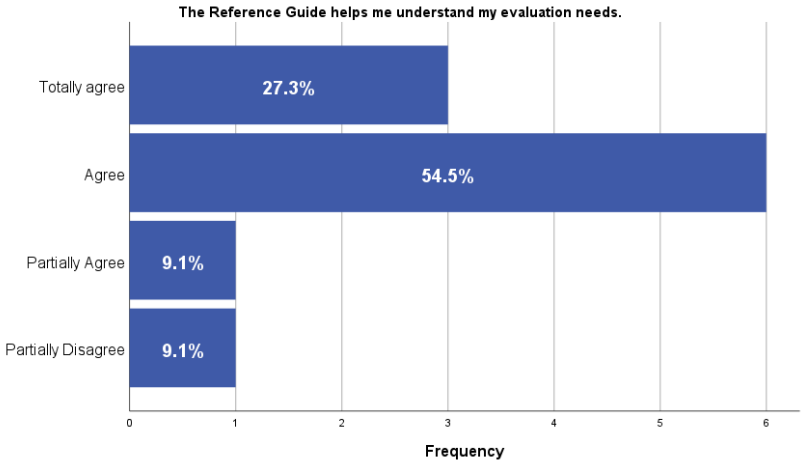


medianRQ2 Usefulness of reference Guide

	N	%
3.00 Partially Disagree	1	9.1%
4.00 Partially Agree	1	9.1%
5.00 Agree	6	54.5%
6.00 Totally agree	3	27.3%

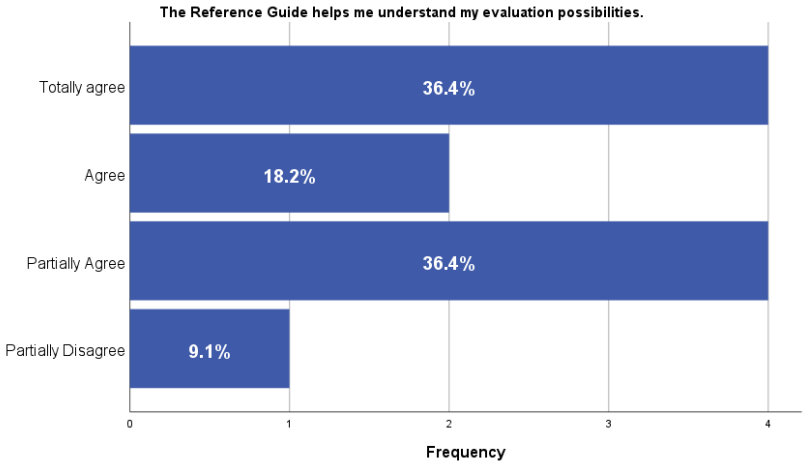
2.1. The Reference Guide helps me understand my evaluation needs.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree

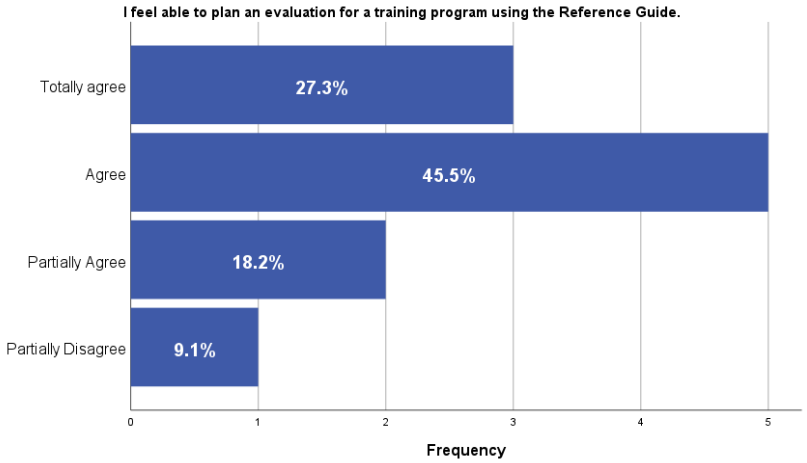


2.2. The Reference Guide helps me understand my evaluation possibilities.

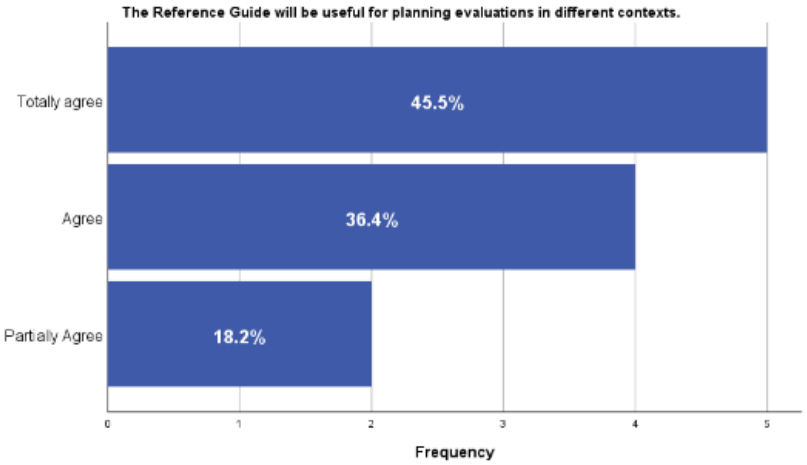
Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



2.3. I feel able to plan an evaluation for a training program using the Reference Guide.
Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



2.4. The Reference Guide will be useful for planning evaluations in different contexts.
Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree

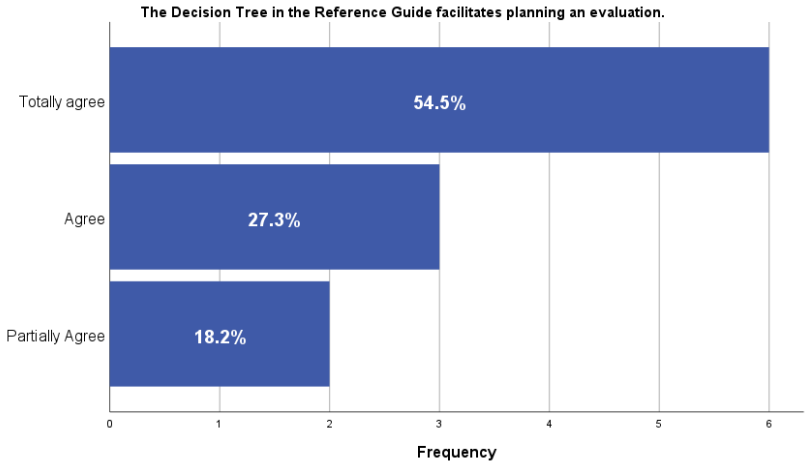


3. Usefulness of the Decision Tree (Q10 – Q 11)
RQ 3: The Decision Tree facilitates designing evaluation plans.

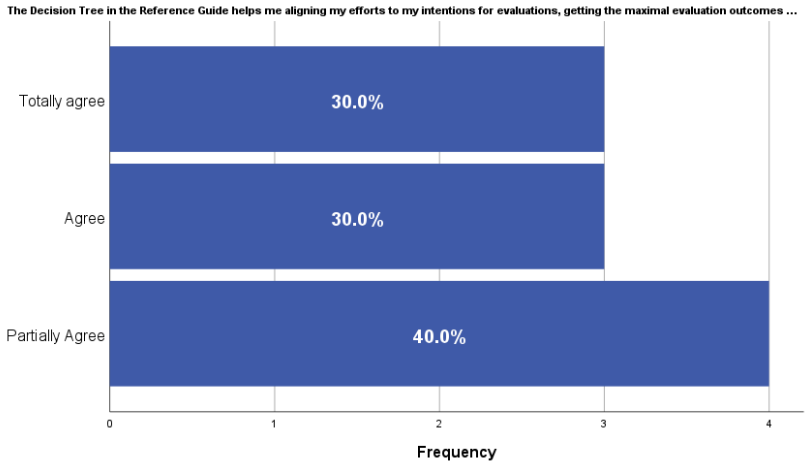
medianRQ3 Usefulness of Decision Tree

	N	%
4.00 Partially Agree	1	9.1%
4.50	3	27.3%
5.00 Agree	2	18.2%
5.50	1	9.1%
6.00 Totally agree	4	36.4%

3.1. The Decision Tree in the Reference Guide facilitates planning an evaluation.
Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



3.2. The Decision Tree in the Reference Guide helps me aligning my efforts to my intentions for evaluations, getting the maximal evaluation outcomes within my possibilities.
Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



4. Usefulness of the Learning Goals Tool (Q12 – Q14)

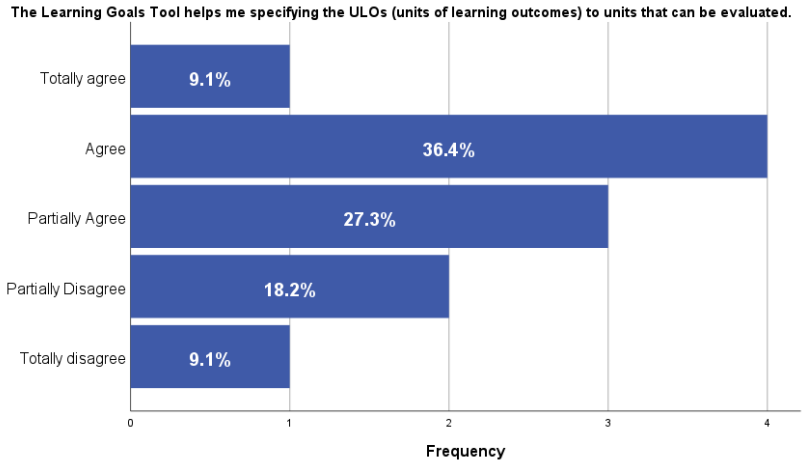
RQ 4: The Learning Goals Tools facilitates designing evaluation plans.

medianRQ4 Usefulness of Learning Goals

	N	%
3.00 Partially Disagree	1	9.1%
4.00 Partially Agree	4	36.4%
5.00 Agree	5	45.5%
6.00 Totally agree	1	9.1%

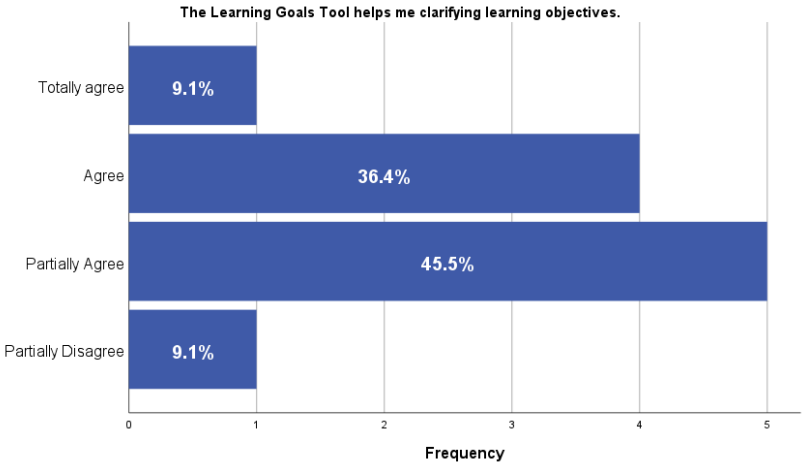
4.1. The Learning Goals Tool helps me specifying the ULOs (*units of learning outcomes*) to units that can be evaluated.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



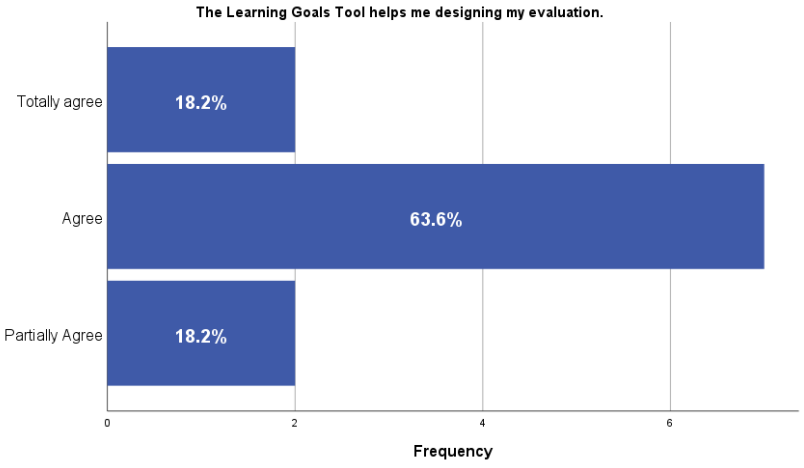
4.2. The Learning Goals Tool helps me clarifying learning objectives.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



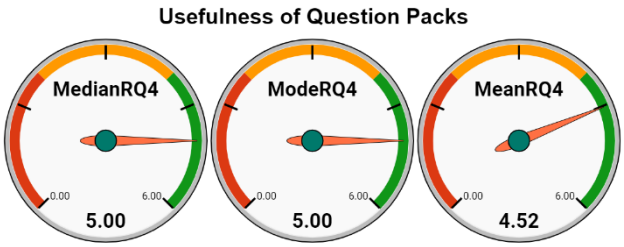
4.3. The Learning Goals Tool helps me designing my evaluation.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



5. Usefulness of the Question Packs

RQ 5: The Question Packs facilitate designing evaluation plans.

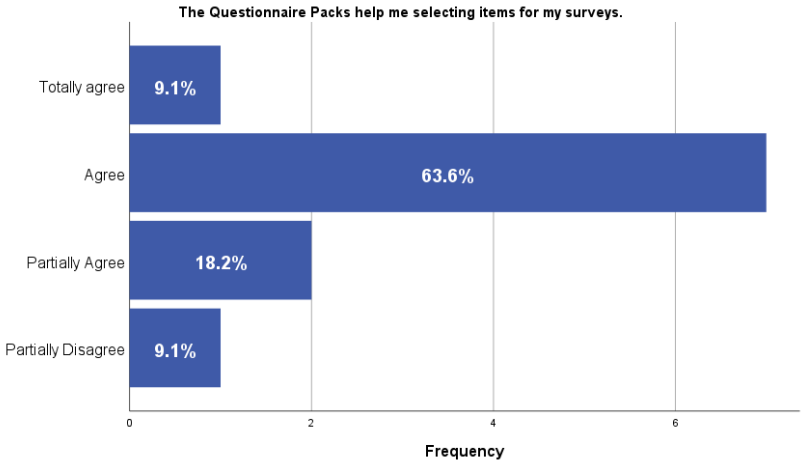


medianRQ5 Usefulness of Question Packs

	N	%
4.50	2	18.2%
5.00 Agree	4	36.4%
5.50	5	45.5%

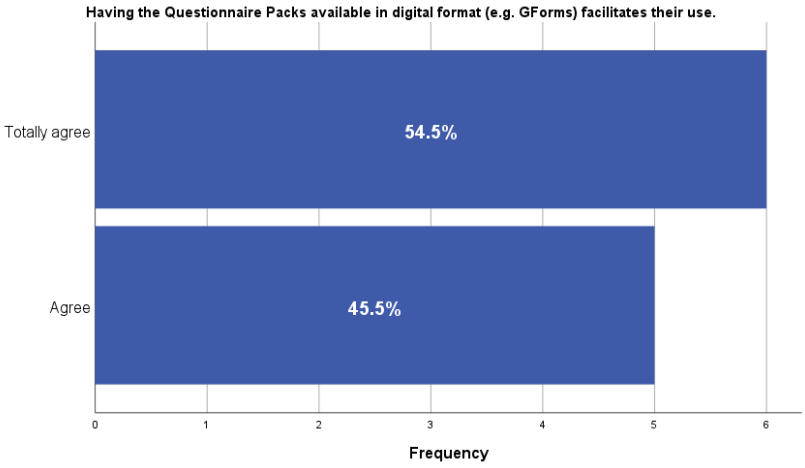
5.1. The Questionnaire Packs help me selecting items for my surveys.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



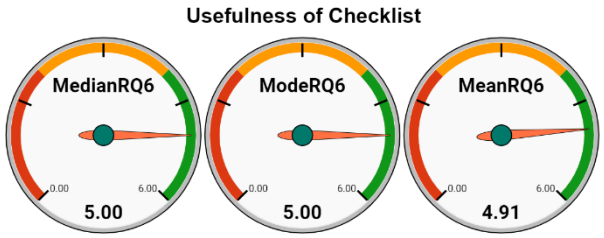
5.2. the Questionnaire Packs available in digital format (e.g., *GForms*) facilitates their use.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



6. Usefulness of the Checklist

RQ 6: The Checklist facilitates designing evaluation plans.

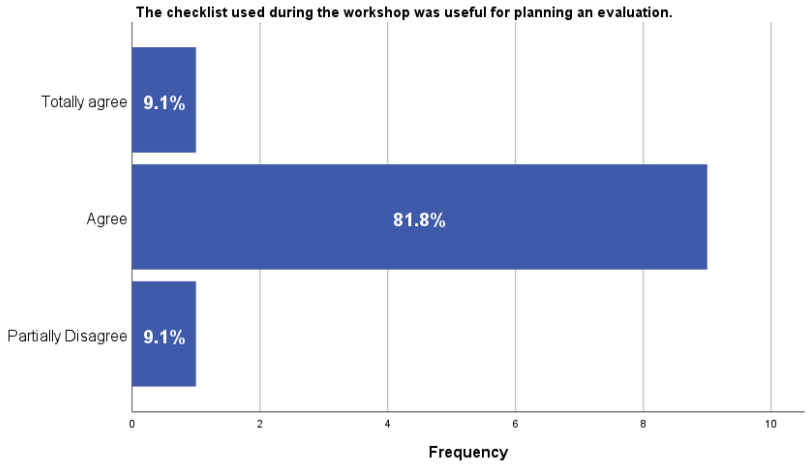


medianRQ6 Usefulness of Checklist

	N	%
3.00 Partially Disagree	1	9.1%
5.00 Agree	9	81.8%
6.00 Totally agree	1	9.1%

6.1. The checklist used during the workshop was useful for planning an evaluation.

Totally Disagree ◯ ◯ ◯ ◯ ◯ ◯ Totally agree



Please answer these questions about the EVALUATION123

18. In the Evaluation123, the levels are independent to each other

☒ True ☐ False ☐ I do not know

19. Learning goals are central to the design of evaluations aiming certifications

☒ True ☐ False ☐ I do not know

20. Pretests are essential conditions for evaluations

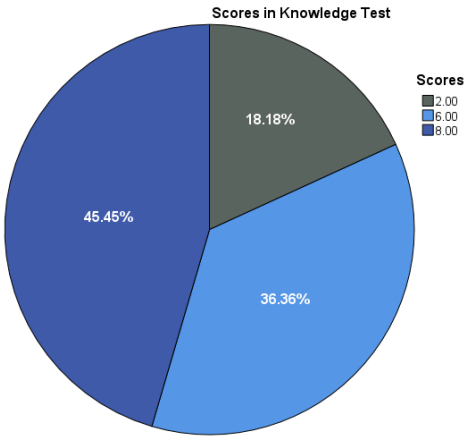
☐ True ☒ False ☐ I do not know

21. Knowledge and skill tests are the only valid ways of assessing learning outcomes.

☐ True ☒ False ☐ I do not know

22. Solid evaluations always contain subjective and objective assessments.

☒ True ☐ False ☐ I do not know



23. Would you like to make any further comment / feedback / suggestion related to the Reference Guide? All inputs are really appreciated.

1. *For quick use, the guide should be translated. In total: it seems [acers]* ***good*** *and compact method*
2. *For persons with no pedagogical background (it) is not so easy to establish the ULO`s*
3. *great work done. How to check > outcome of [scheme] with reality. what is feasible*
4. *Good job! it helps me a lot to design evaluation from the beginning*
5. *very useful tool*
6. *it would be great to have this tool available, especially for small training institutions to help them evaluate effectively*
7. *I wouldn't use the item `level`, which involves a hierarchy, but something like part 1-2-3*
8. *ethics*

UNIVERSITY OF TWENTE.

EVALUATION123

Institution or Organization	**Your name**

Training	
Focus of training (e.g., *upskilling workers in insulation)*	
Target audience: (e.g., *construction workers, plumbers, engineers, architects, customers, etc.)*	*Number of Participants:*

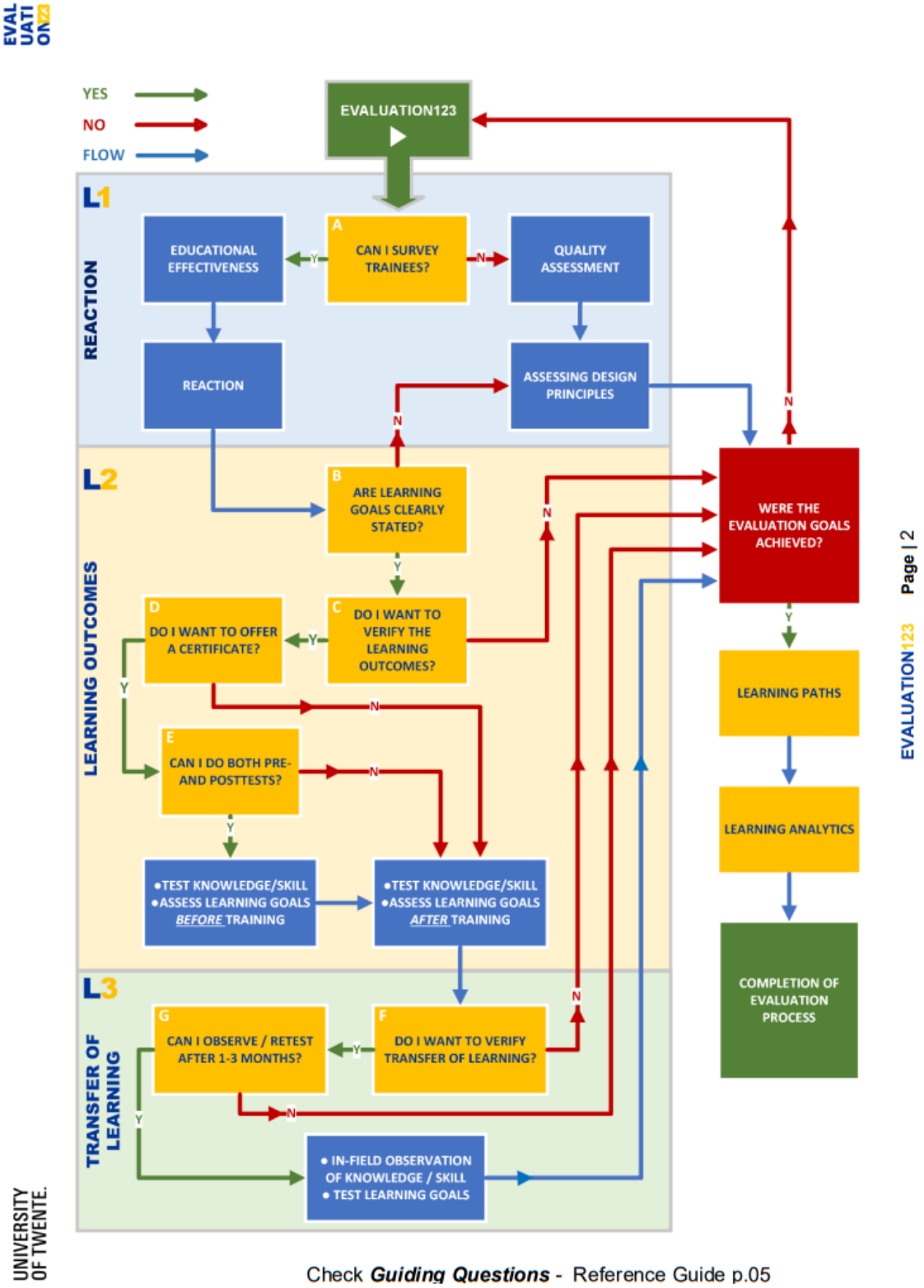
Provider of training	
☐ Own institution or organization	☐ Partner institution or organization
☐ Commissioned training provider	☐ Other:

Type of Training	☐ Voluntary ☐ Mandatory
☐ In-class or on-site ☐ Online ☐ Hybrid ☐ Other: _____	**Duration and length of training** (e.g. *one-hour online training, full-day workshop, part of a modular program, etc.)*

Elements contained in the training program (multiple items may apply)

☐ video lessons	☐ e-learning platform	☐ reading material	☐ Other:
☐ online conferences	☐ hands-on practice	☐ discussion forum	
☐ online tutorial	☐ face-to-face instruction		

Additional information/ comments



Check ***Guiding Questions*** - Reference Guide p.05

LEVEL 1 – REACTION (📖p.7)

Can I survey and/or test trainees? (📖p.3)	
☐ yes In addition to the internal evaluation, you are also able to perform the ***Level 1 - Reaction*** evaluation (in the 💻Question Pack, also in 📖Appendix B). **Plan for Level 1 – Reaction** Check which survey packs you will use for your evaluation: ☐ 1. Mastery experiences ☐ 2. Cognitive development ☐ 3. Teaching ☐ 4. Normative Success ☐ 5. Interaction with others ☐ 6. Fun and enjoyment ☐ 7. Satisfaction with training (So & Brush) ☐ 8. Satisfaction with training (Athiyaman) ☐ 9. E-Learning	☐ no In this case, you can only do an internal evaluation of the quality of the training program. In the Design Principles Checklist (💻Question Pack – ***Level 1 – Quality Assessment***, also in 📖 Appendix A) **Plan for Level 1 – Quality Assessment** ☐ A. Trainee input and involvement ☐ B. Supervisor involvement ☐ C. Trainee attendance policy ☐ D. Trainee preparation ☐ E. Activities during training ☐ F. Supervisor Support ☐ G. Organization Support ☐ H. Evaluation and feedback ☐ I. Accountability ☐ J. E-learning

Are learning goals clearly stated? (📖p.18) Check whether the learning goals are clear. (e.g.: *After the training, participants will be able to **differentiate** types of insulating materials.* First determine the ***level of learning*** (*Conceptual Knowledge, Procedural Skills or Analytical Thinking and then choose **one** action verb*)	
☐ yes You may proceed to Level 2 – Learning Outcomes.	☐ unsure If you want to proceed with any type of test or assessment, you need to first clarify your learning goals (💻Learning Goals Tool, also in 📖p.18)

LEVEL 2 – LEARNING GOALS TOOL (📖 p.19)

UNIVERSITY OF TWENTE. BUS LEAGUE

	Abilities	Description of category
Conceptual Knowledge	**Recognize** and **retrieve** relevant **knowledge**	Demonstrate understanding of facts and ideas by organizing, comparing, translating, interpreting, giving descriptions, and stating main ideas.
	Constructing meaning from instructional messages	
Procedural Skills	**Carry out** or **use** a procedure in a given situation	Solve problems to new situations by applying acquired knowledge, facts, techniques and rules in a different way.
Analytical Thinking	**Breaking down** information into **component parts** and their **relation**	Examine and break information into parts by identifying motives or causes. Make inferences and find evidence to support generalizations.
	Making judgements based on **criteria** and **standards**	Present and defend opinions by making judgments about information, validity of ideas, or quality of work based on a set of criteria.

WHEN	WHO		Level of Learning: _____	WHAT
		will be able to	One Action Verb:	
WHEN	WHO		Level of Learning: _____	WHAT
		will be able to	One Action Verb:	

LEVEL 2 – LEARNING OUTCOMES (📖 p.13)

Do I want to offer certification? (📖 p.15) The need and type of certification will determine which are the different paths you may choose for your evaluation.	
☐ yes **Can you implement both pre-and posttest?** ☐ yes –**Scenario 1**: **Pre-Training-A ▶Training ▶ Post-Training** (📖 p.22) ☐ no – **Scenario 2**: **Training ▶ Post-Training ▶ Pre-Training-B** (📖 p.23)	☐ no You may opt for a one-step evaluation: **Scenario 2**: **Training ▶ Post-Training ▶ Pre-Training-B** (📖 p.23)

Scenario 1: Pre-Training-A ▶Training ▶ Post-Training (📖 p.22)	**Scenario 2: Training ▶ Post-Training ▶ Pre-Training-B** (📖 p.23)
Select the items that you decided to include in your evaluation plan: **Pretest** ☐ Pre-training A – Learning goals ☐ Base knowledge test ☐ Skill test ☐ Skill observation **Posttest** ☐ Post-Training – Learning goals ☐ Knowledge test ☐ Skill test ☐ Skill observation	Select the items that you decided to include in your evaluation plan: ☐ Pre-training B – Learning goals ☐ Post-Training – Learning goals ☐ Knowledge test ☐ Skill test ☐ Skill observation

LEVEL 3 – TRANSFER OF LEARNING (📖 p.25)

Can you perform another evaluation/retest learners after 1-3 months? (📖 p.27) Verifying Transfer of learning usually demands more time, resources and availability from both evaluators and learners.	
☐ yes You might be able to perform a Level 3 evaluation. Proceed to the next section.	☐ no You may reconsider the need for a Level 3 evaluation later, complete the evaluation process with level 2, or further explore the data you have been able to collect so far.

Can you implement an observation test to verify the impact of the training? (📖 p.27)	
☐ yes You may perform another evaluation with your learners. Select the items that you decided to include in your evaluation plan: ☐ Post-training – Learning goals ☐ Observation knowledge test ☐ Observation skill test ☐ Skill observation ☐ Self-Perception Survey (Chauan et al.)	☐ no You may still survey learners. Select the items that you decided to include in your evaluation plan: ☐ Post-training – Learning goals ☐ Self-Perception Survey (Chauan et al.)

EXPLORING PATHS AND DATA (📖 p.31)

Congratulations! In your evaluation design journey, you have made several decisions that will not only measure the quality of your training and its impact, but also provide you with more information on the factors influencing them. This will enable you to take better informed decisions about your training design. Below, you will find two further opportunities for increasing the understanding of your evaluation that you might want to consider.

Learning Paths (📖 p.31)

Do you want to understand the motivation, participation and follow-up plans of your trainees to cater them better?

- ☐ Learning Paths Survey (💻 Learning Paths, also in 📖 Appendix C)

Learning Analytics (📖 p.32)

1. Do you use analytics to understand usage of the digital tool? If yes, which analytics and what insights do they provide you with?

2. What information of learning analytics would you like to obtain by LA? (📖 p.34)
   - ☐ Analyzing learning as an individual process of building up knowledge and skills: i.e., use learning analytics to track the learners' progress, definition of learning goals, implementation of feedback or personalize the recommendation of learning materials.
   - ☐ Analyzing learning as a participative process, in which individuals learn within a community of practitioners: i.e., use learning analytics to identify opportunities to foster participation, build learning groups, or identifying more capable peers and leaders for each group.
   - ☐ Analyzing learning as a collaborative process, focused on advancing knowledge and transforming practices through the interaction amongst individuals and co-creation of materials and artifacts: i.e., use learning analytics to understand the creation process and the relationship between individuals and those materials to better support them.

3. What are your interests and intentions for implementing learning analytics in the future? (📖 p.33)
   - ☐ Integrated learning analytics tools.
   - ☐ External learning analytics tools.
   - ☐ Plugins. *How helpful do you see plugins for your context?*
   - ☐ Spreadsheets (manual control tools). *How helpful do you see the dropout tracking spreadsheet for your context?*
   - ☐ Surveys. *How helpful do you see the provided surveys for your context?*