Production Engineering
https://doi.org/10.1007/s11740-024-01280-4

ORIGINAL PAPER

Learning outcome evaluation in manual assembly

Maria Maier[1] · **Kim Julia Schoenfelder**[1] · **Michael F. Zaeh**[1]

Received: 27 February 2024 / Accepted: 17 April 2024

Abstract
Mass customization and shorter product life cycles are causing ever more variants in production, especially in manual assembly. At the same time, more diverse personnel structures are emerging due to demographic change and labor shortages. This is causing different challenges to production managers, e.g., competence gaps. To meet these challenges, learning in manual assembly becomes increasingly important. The design of the learning process can only be improved by checking whether the processes fulfill their purpose. Various learning evaluation measures are described in general vocational education and competence development, but it is hard to select the right one for the learning process. This paper shows a procedure, how learning evaluation measures can be selected, and how they can measure learning progress. For this, a test person study was conducted to compare different learning evaluation measures and show their usability and advantages in manual assembly. The results support making learning in assembly easier to apply and controllable. In the long term, feeding back the results improves the learning process design.

Keywords Competence development · Manufacturing · Work-based learning · Assembly worker · Learning outcome · Learning evaluation

1 Introduction

Assembly workers face challenges, as the work environment changes due to the greater variety of product variants resulting from mass customization and more frequent product changes due to shorter product lifecycles [1]. Despite the increasing support provided by assistance systems, that support the workers in dealing with complex situations without requiring them to acquire new competencies, new competencies must be learned on a regular basis because using assistance systems is no longer sufficient to handle complex situations completely [2]. As a consequence of demographic change, the shortage of skilled workers has created an additional need for competencies acquisition, as companies are forced to adapt their workforces and cannot hire (skilled) workers as they need [3]. As a result, more and more semi-skilled and unskilled workers are being empowered for more complex tasks [4]. Depending on the complexity of the product and assembly processes and the amount of assembly time for one product, learning can take time from a few days up to months [5]. As learning demands different, additional resources (e.g., experts, extra room), it is necessary to control the learning process [6].

Learning is always associated with a goal and a schedule for achieving the goal [6]. To verify the achievement of the goal, an evaluation of competence or learning outcome is required [7]. The measurement should be integrated into the workplace equipment or the learning process for increased efficiency. Therefore, it is necessary to select a measurement method that fits the circumstances of the workplace and the learning process. In that regard, this paper introduces different evaluation measures, describes a procedure for situation-specific selection, and evaluates the measures in the context of a study.

The paper starts with an overview of evaluation measures and a summary of previous research in this area in Sect. 2. Section 3 presents the procedure for workplace-specific selection of the measurement procedure. Afterwards, the application of selected measures in a study is described and verified based on a manual assembly use case, which is presented in Sect. 4. Lastly, the conclusion and further research perspectives can be found in Sect. 5.

✉ Maria Maier
maria.maier@iwb.tum.de

[1] Institute for Machine Tools and Industrial Management (iwb), Technical University of Munich, Boltzmannstrasse 15, 85748 Garching, Germany

Published online: 06 May 2024

2 State of the art

2.1 Measures for evaluating competence and learning outcome

When researching evaluation measures, it was noticeable that there are two groups of measures for evaluating competence. On the one hand, measures have been developed to measure competence holistically. These are primarily complex measures carried out by trained or experienced personnel. These measures are called measures for competence evaluation (CE) in the following. On the other hand, some measures do not claim to represent competence comprehensively but rely on the fact that artifacts can partially define competence, e.g., required execution time and number of mistakes [8]. These measures are called measures for learning outcome evaluation (LOE) because they can map learning progress, but no long-term determination of competence is possible [7]. The measures can be further differentiated into qualitative or quantitative and subjective or objective measures, which are relevant for the implementation and significance.

Qualitative measurement methods can be used to assess the type and quality of formally, non-formally, and informally acquired competencies. The measurement is holistic and relates to the meaning and interrelationships of competencies. Quantitative measurement methods are based on the assumption that competencies are measurable and scalable [9]. Subjective measures are based on self-assessment or assessment by others. Objective measures are based on data collected externally by technologies or people in a standardized way.

A literature review was conducted based on the information about learning evaluation given in [10–15]. For further queries in the databases 'Scopus' and 'Google Scholar', the following terms were used: 'measurement of learning success', 'assessment of learning progress', 'performance evaluation', 'knowledge assessment', 'measurement of learning gains', '(methods of) competence assessment', 'competence development in assembly', 'assessment of competence'. The results were sorted out, if they do not include relevant competencies, especially technical and methodological competencies. Additionally, the suitability for the work context, especially for the industrial and the assembly context was evaluated based on the measures' descriptions. The remaining measures were categorized as CE or LOE measures according to the definition given above.

Based on the literature review, eleven CE measures and 15 LOE measures have been selected as examples to be used with the procedure developed in this paper to make human learning in assembly quantifiable. The CE measures are shown in Table 1 and the LOE measures are summarized in Table 2.

2.2 Similar approaches

As shown in Sect. 2.1, different evaluation measures are described in the literature. They are characterized by using general aspects of learning, and not all of them were tested in manual assembly but in other production near-production environments in companies. Therefore, an additional literature search was conducted to gain an overview of further aspects relevant to the manual assembly learning evaluation.

Arena et al. [20] followed the work of Perini et al. [21] and formulated an ontology for training evaluation for a trainee group. They summarize relevant criteria to evaluate training, e.g., Kirkpatrick competence level and training KPIs. Kuna et al. [22] gave an overview of which factors are essential for a holistic review of industrial training. The evaluation included the trainees' performance and an assessment

Table 1 Measures for competence evaluation

Short Cut	Measure	References	Property
CE1	ASSESS	[10]	Quantitative, subjective
CE2	Persolog	[11]	Quantitative, subjective
CE3	KODE procedure	[12]	Quantitative, subjective, objectifying
CE4	Learning part (German: LERNSTUECK)	[13]	Qualitative, subjective
CE5	COMPRO+	[14]	Quantitative, subjective
CE6	Competence wheel and competence matrix	[9]	Quantitative, subjective
CE7	BIP	[15]	Quantitative, subjective
CE8	Procedure according to Hertle	[16]	Quantitative and qualitative, subjective
CE9	Video analysis	[17]	Quantitative and qualitative, subjective
CE10	Competence pass	[18]	Quantitative, subjective
CE11	Methodology for multivariate measurement of technical-methodical competencies for the production	[19]	Quantitative, subjective

Table 2 Measures for learning outcome evaluation

Short cut	Measure	References	Property
LOE1	Free impression description	[25]	Qualitative, subjective
LOE2	Processing of statement lists	[25]	Quantitative, subjective
LOE3	Ranking measures	[25]	Qualitative, subjective
LOE4	Ratings	[26]	Quantitative, subjective
LOE5	Behaviorally anchored rating measures	[27]	Quantitative, subjective
LOE6	Task- and goal-oriented assessment measures	[25]	Quantitative, subjective
LOE7	External observation	[28]	Quantitative or qualitative, subjective
LOE8	Automatic time measurement	[24]	Quantitative, objective
LOE9	Manual time measurement	[29]	Quantitative, objective
LOE10	Measurement of action indicators	[19]	Quantitative, objective
LOE11	Test with open questions	[30]	Quantitative, mostly objective
LOE12	Multiple-/Single-Choice Test	[31]	Quantitative, objective
LOE13	Test with false sentences and gap text	[30]	Quantitative, objective
LOE14	Roleplay	[32]	Qualitative, subjective
LOE15	Web of knowledge	[30]	Quantitative, subjective

of the training need, the training design, and the trainer's performance. These research approaches give an overview of relevant frameworks for implementing learning evaluation. Contrarily, in Sehr et al. [8], Hertle et al. [16], Glass and Metternich [19], Gross et al. [23], and Wilschut et al. [24], different learning outcome evaluation systems are discussed, and challenges are identified, e.g., focus on learning transfer, compatibility with digital learning systems.

The reviewed literature has shown that general models explain the learning process and the associated evaluation process. Many different measurement methods were developed in production or its environment. They have varying requirements and show quality criteria. In summary, an open point in research is to provide a simple approach to select an evaluation measure for industrial applications. Another point is to demonstrate the applicability and validity of measures that have not been widely used in manual assembly so far.

3 Procedure for selecting the CE and LOE measures

The first result of the research is a procedure for selecting suitable CE and LOE measures for the learning process. Each CE and LOE measure presented in Sect. 2 supports learning control differently and places additional requirements on the work environment. To compare the methods with each other, a two-step procedure was developed, which helps to select the suitable measures for the targeted workplace.

In step 1, the methods are assessed based on the five following evaluation criteria.

- *K1: Quality criteria (objectivity, reliability, and validity)*
  Objectivity defines the independence of the results of external influences, like the executing person or the surroundings. Reliability is the extent to which repeated measurements of a test object lead to the same effect. Validity describes the degree of accuracy with which a method measures the property for which the measurement is performed.
- *K2: Time requirements*
  Time requirements label the duration and frequency of implementation, execution, and evaluation.
- *K3: Technical requirements*
  The necessity of technological tools and their complexity is summarized as technical requirements.
- *K4: Physical requirements*
  Physical requirements describe the need for extra space caused by, e.g., space for extra sensors or people, which need to be placed near the learning person for observation, and restrictions regarding the workplace design are applicable.
- *K5: Personnel requirements*
  Personnel requirements include the expenses for additional staff and their qualification needed for implementation, execution, and evaluation.

The measures concerning the evaluation criteria are assessed using a scale of 5—'Very good' to 0—'Unsatisfactory' with an even number to force a clear decision rather than deciding for the middle [33]. The description of the criteria is given in Table 3.

The evaluations of the measures of Sect. 2 were carried out by the authors based on the literature and are presented in Tables 4 and 5. The remaining measures are evaluated in

Table 3 Overview of the criteria assessment description

	0	1	2	3	4	5
K1: Quality criteria	Insufficient quality for all three quality criteria	Insufficient for at least two of three quality criteria	Insufficient for at least one of three quality criteria and a second without high quality	Insufficient for at least one of three quality criteria	One or two criteria without high quality	High quality for all three quality criteria
K2: Time requirements	Multiple implementations, time-consuming execution and evaluation	Implementation and evaluation with high effort, execution with low or medium effort	Implementation with high effort, execution, and evaluation with medium effort	Implementation with low effort, execution, or evaluation with high effort or both with medium effort	Implementation with high effort, execution and evaluation with low effort	One-time implementation, quick execution and evaluation
K3: Technical requirements	Additional expensive technologies, equipment, or materials	Additional medium-cost technology or equipment, e.g. new software on existing computer	Additional digital, cheap technology or equipment, e.g. stopwatch	Additional analog material for execution	Digital low-effort material, e.g. questionnaire	No particular technologies or materials for implementation, execution and evaluation
K4: Physical requirements	A lot of extra space	Additional space for 2—3 person and for extra technology	Additional space for 2—3 people	Additional space for employee for test execution with small technology	Additional space for one person through execution or a small technology	No additional space for implementation, execution, and evaluation
K5: Personnel requirements	Several employees with special qualifications	2—3 employees once for implementation, several employees with knowledge on work task, employee, and evaluation for each time of execution and evaluation	2—3 employees once for implementation, one employee with knowledge on work task, and employee, for each time of execution and evaluation	One employee once for implementation, one employee with knowledge on work task, and employee, for each time of execution and evaluation	One employee once for implementation, no employee for execution, one employee with knowledge on work task, and employee, for each time of evaluation	No other staff for implementation, execution, and evaluation

Table 4 Criteria evaluation for the measures of competence evaluation

Measure	K1	K2	K3	K4	K5
CE1	5	3	5	3	3
CE2	5	5	4	4	2
CE3	5	4	4	3	1
CE4	4	1	5	3	2
CE5	4	2	5	3	3
CE6	X	2	4	5	4
CE7	4	3	4	5	3
CE8	X	2	3	3	3
CE9	X	1	2	2	2
CE10	X	4	5	5	4
CE11	4	2	4	5	3

Table 5 Criteria evaluation for the measures of learning outcome evaluation

Measure	K1	K2	K3	K4	K5
LOE1	1	3	5	4	3
LOE2	4	3	5	4	3
LOE3	X	4	5	5	3
LOE4	3	5	5	5	3
LOE5	4	2	3	3	2
LOE6	3	2	5	5	4
LOE7	4	3	5	4	3
LOE8	4	5	3	4	4
LOE9	4	3	4	4	3
LOE10	4	3	3	4	4
LOE11	3	4	5	4	4
LOE12	4	5	4	4	5
LOE13	3	4	5	4	4
LOE14	X	2	3	3	3
LOE15	X	4	5	4	4

comparison to this measure and under consideration of the descriptions and the literature given in Tables 1 and 2. No rating, represented by 'X', is assigned if a criterion cannot be evaluated for a measure due to a lack of information in the literature.

In step two, the user defines which criterion may be applied and to what extent by taking the learning process and the workplace into account. The user could be the team leader of the trainee or the person preparing the learning process. The decision follows the same description as the literature-based assessment for the measures, shown in Table 3. If, for example, there is no possibility of additional space for implementing the measure, the user selects '5'. Measures with lower ratings than those selected by the user are then excluded. Thus, checking each literature-based measure evaluation filters which measures are possible. In the end, a short list with measures fitting the user requirements is generated. The short list is then checked to ensure that the measures' goals align with the required goal, e.g., assessing each person's learning outcome individually or comparing employees to each other.

4 Study to test CE and LOE measures in manual assembly

As part of the study, the procedure of Sect. 3 was applied to a workstation in the learning factory of *iwb*. The workstation is presented in Sect. 4.1, and the procedure is applied in Sect. 4.2. The focus of the selection and the study was to check if measures not regularly used in assembly are suitable for an assembly learning process. For this, a preliminary study was executed to validate the learning supporting assistance system, summarized in Sect. 4.3, followed by the main study, described in Sect. 4.4.

4.1 Workstation

An assembly workstation in the learning factory of the *iwb* was selected as the test workstation. This was a standing workstation, supplemented by a shelf for material provision, as depicted in Fig. 1. All required materials and tools were provided directly at the workstation at fixed places. The people assembled a three-stage planetary gearbox at the workstation with the help of step-by-step instructions on a tablet. No prior knowledge was required for the assembly task. The necessary assembly steps, the required processes and technical expertise were identified by a thorough work analysis. The competencies required for assembly were, thus, derived. The task was divided into the assembly of

Fig. 1 Workplace for the study

individual gear stages, the gear stages' composition, and the engine's assembly.

4.2 Applied procedure to select measures

For applying step one of the procedure, the measures are reviewed, and their assessment, shown in Tables 4 and 5, is used for the review. For step two of the procedure, the requirements were defined:

- *K1: Quality criteria (objectivity, reliability, and validity)*
  Following the study's objective, it was necessary to ensure high levels of objectivity and reliability, as well as the best possible validity. As a result, a rating of 4 or 5 is considered sufficient.
- *K2: Time requirements*
  The study was conducted to closely replicate real-world circumstances. It was assumed that time plays a crucial role, but preparation and implementation time are not critical. A range of 3 to 5 was selected.
- *K3: Technical requirements*
  The learning measures, implemented at the examplary workplace, require a tablet system with pre-installed software. This technical equipment can be used by the CE or LOE, but more complex technical equipment should be avoided. Therefore, the K3 is chosen from the range of 3 to 5.
- *K4: Physical requirements*
  The study was conducted at the *iwb* learning factory, where space and physical requirements were not a concern. The rating scale ranged from 0 to 5.
- *K5: Personnel requirements*
  The study's instructor can implement measures for CE or LOE but has no additional qualifications for special measures. The range of 3 to 5 was chosen.

When comparing the requirements with the assessment in Table 4, it became clear that CE1, CE7 and CE10 are possible CE measures. Looking at the goals of the measures, only the competence pass (CE10) is suitable for the study. ASSESS (CE1) and BIP (CE7) measures assess personality characteristics, which are not applicable to the study's objective of evaluating the employee's competence.

Comparing the requirements with the assessment of Table 4, LOE2, LOE3, LOE7, LOE8, LOE9, LOE 10, LOE12, LOE15 are possible LOE measures. Checking the goal of the measure, LOE3 and LOE15 are sorted out. LOE3 compares the results of different employees. This is not in line with the objective to get quantitative results. LOE15 focuses on feeding back the learning progress to the teacher which is not the goal of the study. Processing of statement lists (LOE2) and Multiple-/Single-Choice-Test (LOE12) are focusing on the same goal of assessing knowledge. The Single-Choice-Test (LOE12) was selected because it is easier to understand by the subject. Both automatic time measurement (LOE8) and manual time measurement (LOE9) measure time and heavily overlap. Consequently, only one of these measures had to be selected. The preliminary study focused on the functionality of the learning measure. Therefore, LOE9 was used to minimize additional implementation effort. In the main study, LOE8 was used instead. The observation (LOE7) can integrate the measurement of action indicators (LOE10), combining them as an observation with a focus on mistakes and instruction dependency.

A total of one CE measure and four LOE measures were used in the study.

4.3 Preliminary study

4.3.1 Aim of the study

The preliminary study aimed to test if the prototypical assistance system implemented on a tablet supports the learning process. Additionally, knowledge about the practicability of the study progress should be gained.

4.3.2 Workplace and work activity

The study was conducted at the assembly workstation described in Sect. 4.1. At the workstation, the subjects had to follow the assistance system's instructions during each assembly run, but they could deviate from the specified assembly sequence after the first run if they found a more convenient way. The available instructions were based on three competence levels in the preliminary study. For the first run, the test persons got detailed information by provided videos, pictures, and text. If the assembly run was completed without mistakes, a leaner instruction was given, containing only images and text. If that was not the case, the person continued with the detailed one until she or he assembled it without mistakes. If the test person completed the task without mistakes and within a target duration, the instruction changed to an information slide with the information of the necessary material without step-by-step instructions.

4.3.3 Preliminary study procedure

The study procedure followed a guideline prepared in advance. This ensured that the process was the same for all subjects and minimized influence of experimental aspects and other disruptive influences. Initially, the subjects received information about the study and data processing, for which they signed an informed consent. Subsequently, the subjects' demographic data were collected using a questionnaire.

The test persons were informed that the target is to assemble without mistakes in a predefined time of 285 s. The number of assembly runs was not fixed. The test persons repeated the assembly task as long as they needed more than the predefined time. According to this aim, the assembly time per run was measured manually, and the study leader observed the test persons and checked the finalized products to count the runs needed for a zero-defect part. The test persons were also asked to fill out NASA-TLX (NASA Task Load Index, [34]) and SUS (System Usability Score, [35]) questionnaires to assess the assistance system's instruction quality. The NASA-TLX evaluates the mental, physical, and temporal stress of the person. With this information, conclusions on the learning process and the support of the system are possible. The SUS enables the assessment of the system's user friendliness. To overview the whole process, the NASA-TLX and the SUS were filled out each time after the person completed one goal and got a new instruction. After the assembly run, a final interview was carried out to get qualitative feedback for the assistance system.

4.3.4 Findings of the preliminary study

The preliminary study was executed with 18 test persons. Eleven subjects were male, and seven were female. Four subjects stated that they had previous experience in the assembly field.

The subjects needed three to six assembly runs to achieve the targeted time. Through runs one to four (depending on the maximum runs), the mistakes (deficiencies of the end product) were reduced to zero, and the time was constantly reduced from a maximum of 22 min to a minimum of nearly three minutes. These measures showed that learning progress appeared, and the learning curve was as described by, e.g., [36]. The NASA-TLX resulted between 25 and 42 in average, which indicates that subjects perceived low to medium stress level. Especially, the information reduction between the first and second instruction showed an increase in the NASA-TLX results. The SUS showed an average result of 81 for the first instruction, 80 for the second instruction, and 78 for the third instruction. Respondents indicated that the last one contained too much information for their learning level. Overall, the results show a good acceptance and usability, according to Bangor, Kortum and Miller [37].

Analyzing the functionality of the assistance system and the study process, conclusions were drawn for the construction of the main study. First, even if the study showed that dynamical instructions are usable, it did not support assessing the usability of the CE and LOE measures because the comparison between the assembly runs is not possible due to the change of instructions. Second, the test persons rated a high SUS for the assistance system but mentioned a technical issue during the interview. The loading time for the images was too long. This was caused by the online system that retrieves images from a server. The image integration was modified for the main study, and the loading time was reduced to less than one second. Lastly, the study process was evaluated, and it was decided to use a similar one with more CE and LOE measures in the main study. Instead of structuring the study by targeted time and zero mistakes, the main study was conducted by running the assembly five times, which is the number of runs most test persons in the preliminary study needed to reach the targeted time.

4.4 Main study

4.4.1 Aim of the study

The purpose of the study was to test the selected CE and LOE measures. It focused on the validity of the measures with respect to the learning progress of the test subjects and the consistency of the measures. The measures were also assessed in terms of their comprehensibility, manageability, and acceptance.

4.4.2 Workplace and work activity

The workplace remained the same as during the preliminary study. The instructions given in the assembly runs were changed to a static instruction with the same amount of information throughout the five assembly runs. The test persons received instructions with pictures and text through all five runs.

4.4.3 Study process and application of the measures

The study process followed a guideline prepared in advance, which started in the same way as the preliminary study.

The selected measures were adapted to the workplace and the assembly activity. The self-assessment of the subjects' competencies was conducted using a questionnaire. In this questionnaire, the test subjects were asked to rate themselves concerning various statements on a scale from 1—"Does not apply" to 7—"Fully applies". An extract of the questionnaire is shown in Fig. 2. The competencies assessed were subdivided into the areas of technical and methodological competencies, assessment ability, and follow-up awareness. The subjects performed this self-assessment before the first assembly run.

During the assembly runs, the assembly times of the subjects were registered automatically by the assistance system. The time measurement started when the instructions began and ended when the last instruction was closed. The subjects were informed about the time measurement beforehand but not about their assembly time during the study.

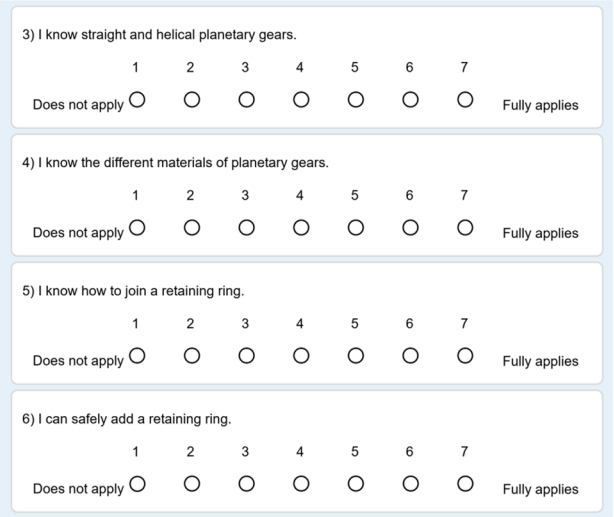

Fig. 2 Self-assessment questionnaire (extract)

Based on a previously prepared observation sheet, the study instructor observed the subjects during the assembly runs. The observation sheet was filled out for each assembly run. On this sheet, a seven-point Likert scale was used to assess the extent to which the subjects acted calmly in general and in difficult situations. Furthermore, the degree of instruction dependency was recorded for each assembly run using a Likert scale. The numbers of questions asked, corrections made, and errors made were also noted on the observation sheet.

After the assembly runs, the subjects assessed their competencies again and answered the single-choice test on the work task. This was done on a tablet at a place different from the assembly workplace to avoid influencing the subjects. The single-choice test assessed the subjects' acquired knowledge. The test consisted of ten questions with several answer options, each with one correct answer. Each correct answer was awarded one point. The questions were related to individual components, the assembly sequence, and the activities to be performed.

The final interview was used to determine which measures were accepted by the subjects and to what extent. Furthermore, the test subjects assessed the comprehensibility and manageability of the measures. For the interview, a guideline with questions and a ten-point answer scale was prepared in advance, which was filled out by the test administration.

4.5 Results

4.5.1 Description of the sample

33 subjects participated in the study. 20 subjects were male, and 13 were female. The subjects in the study were between 20 and 50 years old. Almost 70% of the subjects were students at the time of the study. The remaining subjects held occupations in a variety of industries. 30% of the subjects stated that they had previous experience in the assembly field. The subjects were recruited at the Technical University of Munich (TUM) and among acquaintances. A prerequisite for participation in the study was that the subjects did not know the assembly object beforehand. After the initial review of the collected data, the data set of one test subject was excluded from further evaluation because the data differed from the other data sets, and the person reported physical discomfort during the study. These were considered outliers to avoid distorting the review, and the evaluation was carried out with the remaining 32 data sets.

4.5.2 Evaluation of the measures

The study was evaluated in a closed manner after the participation of all subjects. Descriptive analysis was used to describe the data with key figures (mean value and standard deviation) and graphs. An inferential statistical analysis with statistical tests was carried out to investigate the research goal. The aim was to conclude unknown parameters of the population based on the known parameters of the sample. The statistical analysis of the data was carried out using Excel and the statistical programs R and JASP. Unless otherwise stated, a significance level of $\alpha = 0.05$ was used for all statistical calculations.

4.5.2.1 Self-assessment The statistical evaluation of self-assessment based on the competence pass (CE10) was carried out by three directed t-tests for the competence areas of technical and methodological competence, assessment ability, and follow-up awareness. For this purpose, each subject's average score in the three competence areas was first calculated. The values for technical and methodological competence can be seen in Fig. 3.

For technical and methodological competence, the following hypotheses resulted from the directed t-test:

- H_0: The technical and methodological competence score before assembly is higher than or equal to the score of the technical and methodological competence after assembly.
- H_1: The score of technical and methodological competence after assembly is higher than the score of the technical and methodological competence before the assembly.

The hypotheses for assessment ability and follow-up awareness were formulated analogously.

The evaluation of the t-tests showed a significant result for each of the three competence areas, which means that the respective hypotheses H_0 were rejected. The technical

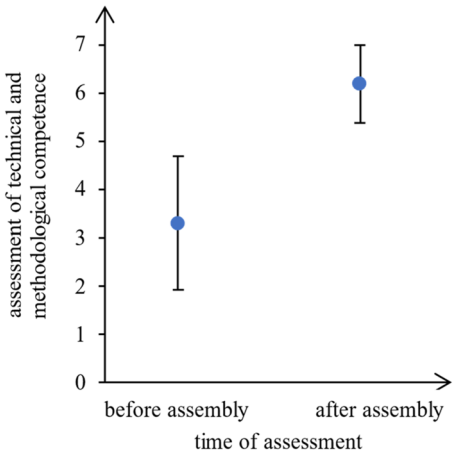

Fig. 3 Assessment of technical and methodological competence

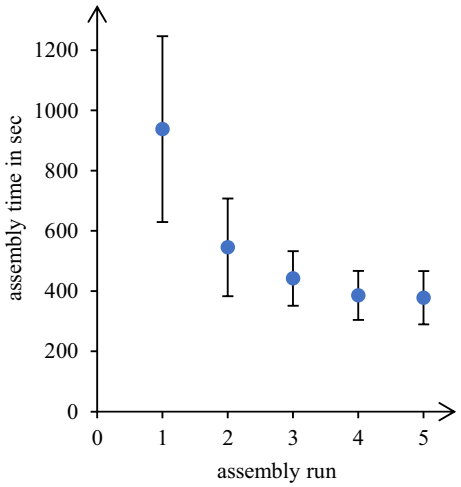

Fig. 4 Time measurement

and methodological competence with the test parameter t value $t(31) = -12.82$ and p value $p < 0.001$ was significantly higher after the five assembly runs than before.

Similarly, assessment ability ($t(31) = -9.64$, $p < 0.001$) and follow-up awareness ($t(31) = -6.92$, $p < 0.001$) were significantly higher after the assembly runs. The effect size Cohen's d, according to [38], indicated a large effect for all three competence areas ($d_1 = 2.27$, $d_2 = 1.70$, $d_3 = 1.22$).

4.5.2.2 Time measurement The automatic time measurement (LOE8) was statistically analyzed using a univariate single-factor analysis of variance with repeated measures. This was used to check whether there were statistically significant differences between the assembly times of the assembly runs. The dependent variable was the assembly time of the test subjects, and the independent variable was the number of assembly runs. The results are shown in Fig. 4.

Due to technically missed data points, the data sets of two subjects were excluded from the statistical evaluation, so the analysis of variance was performed with 30 data sets. The hypotheses of the analysis of variance were:

- H_0: The assembly time is constant for all five assembly passes, which means that the number of assemblies performed has no influence on the assembly time.
- H_1: At least the assembly times of two assembly passes differ. Thus, the number of assemblies already performed has an influence on the assembly time.

After checking the preconditions, the analysis of variance became significant with the test parameter F (1.64, 47.56) = 87.72 and $p < 0.001$, so the hypothesis H0 was rejected. A post hoc test with Bonferroni significance level correction was then performed to determine which assembly runs differed significantly. A significant difference was thus found between the first and all other assembly runs, the second and fourth assembly runs, and the second and fifth assembly runs. To assess how high the statistically significant effect is, the effect size f was calculated using the effect size measure η^2. With $\eta^2 = 0.75$ and an effect size of $f = 1.13$, there was a large difference between the assembly times according to Cohen's effect size limits [38].

4.5.2.3 Observation During the observation (LOE7), evaluating the instruction dependency was particularly interesting. The aim was to investigate whether the degree of instruction dependency differed significantly in the five assembly runs. Since the available data did not meet the requirements for variance analysis, the nonparametric Friedman test was conducted to examine central tendencies in a sample. The hypotheses for the degree of instruction dependency were:

- H_0: Central tendencies of at least two of the five observations of the degree of instruction dependency do not differ.
- H_1: Central tendencies of at least two of the five observations of the degree of instruction dependency differ.

The Friedman test was significant with the test parameter chi-squared $\chi^2(4) = 101.84$ and p value $p < 0.001$, so the null hypothesis was rejected. Thus, at least two observations for the degree of instruction dependency have differing central tendencies. A pairwise comparison was used to determine which assembly runs had significant differences. The progression of the degree of instruction dependency can be seen in Fig. 5.

Of particular noteworthiness are the significant differences between the first and third, second and fourth, and third and fifth assembly runs. This underlines the apparent decrease in the use of the assembly instructions. For

the effect size, the value was $w = 1.78$. A strong effect was present according to the limits for a small, moderate, and strong effect.

4.5.2.4 Single-choice test The single-choice test (LOE12) was evaluated only descriptively. The average score of the subjects was 8.59 points, with a standard deviation of 1.22 points. The minimum score was 6 points, and the maximum was 10 points.

4.5.3 Qualitative evaluation

After the separate evaluation of the measures, a combined assessment was performed, as well as the review of the interview.

4.5.3.1 Multiple regression and qualitative conclusions Following the separate evaluations of the measures, it was examined which of the recorded values could be used together to predict the assembly time of the fifth assembly run. It was assumed that the assembly time, as a result of an objective measure, would best reflect the competence and the learning progress. Additionally, instead of a competence assessment, assembly time is a value that is widely used in manual assembly to evaluate learning outcomes. Therefore, the multiple regression aimed to test how the other measures can be used together to get a conclusion on the learning progress against the industrial standard of assembly time. Multiple regression was performed using assembly time as the criterion and various predictors from the measures enacted. Some predictors showed a significant difference between the assembly runs in the separate analyses. It should be noted that the analysis did not intend to create a specific predictive model. This is not possible due to the limited number of test subjects. The goal of running this multiple regression was to determine which variables in combination are suitable for predicting assembly time and to discuss which variables should be used together to get valuable information about the learning outcome. The predictors were selected first stepwise and then in a theory-driven way [33]. The goal was to obtain models in which all model predictors become significant, and the model has the highest possible explanation of variance.

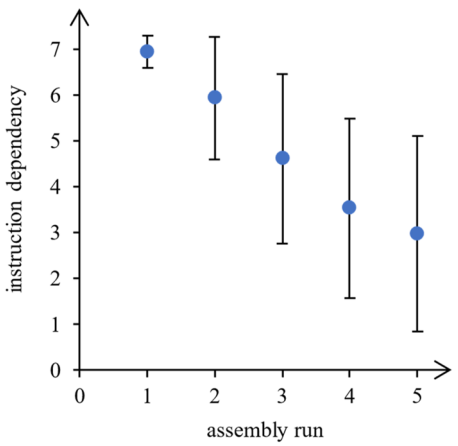

Fig. 5 Instruction dependency

As a result, a model with the predictors 'competence self-assessment before assembly runs', 'competence self-assessment after assembly runs', 'difference in competence assessment', 'degree of instruction dependency in the fifth assembly run', and 'assembly time in the first assembly run' emerged. This model was significant at $p = 0.001$ and had an explanation of variance elucidation of $R^2 = 43.39$ (possible maximum of $R^2 = 100$).

The multiple regression showed that single predictors such as instruction dependency ($R^2 = 14.33$) or single-choice test scores ($R^2 = 16.31$) can also predict assembly time. However, these predictors only have a low explanation of variance. Therefore, it is helpful to use models with multiple predictors that use values from the self-assessment of competencies, the observation, and the time measurement. These have a significantly higher explanation of variance. Utilizing a multivariate approach that integrates both objective data and subjective measures for predicting assembly times offers a more comprehensive and accurate method, by capturing a broader spectrum of influencing factors. The objective measures quantitatively assess the learning progress, and the subjective measures quantify the experience gained by the learner through the learning process. This combines the internal and external view on the learning progress which is promising to give the best outlook on the long-term learning result.

4.5.3.2 Evaluation of the interview The interview aimed to evaluate the measures used in the study regarding their degree of reflecting learning outcome and competencies, comprehensibility, and acceptance. Each question was assessed on a scale from 1—'Not at all' to 10—'Completely'. The results are summarized in Table 6.

Subjects rated all measures as reflecting learning outcome and competencies, but observation was seen as the best and Single-Choice Test as the worst. Additionally, the comprehensibility of all applied measures was rated as very high. The single-choice test had the highest score and observation the lowest.

The acceptance of the single-choice test was rated highest, and the acceptance of the time measurement was rated lowest.

In addition, for the measures carried out during the learning process (time measurement and observation), subjects were asked to rate the impact of the measures on themselves. A two-tailed t-test indicated that time

Table 6 Mean and standard deviation (in brackets) of the rating of the measures in the interview

Measure	Reflection of learning outcome and competencies	Comprehensibility	Acceptance	Impact	Manageability
Competence self-assessment	7.97 (2.02)	8.88 (1.39)	7.47 (2.42)	–	9.03 (1.49)
Time measurement	7.97 (1.20)	8.91 (1.44)	5.47 (3.12)	4.72 (2.39)	–
Observation	8.09 (1.99)	8.78 (1.66)	6.97 (2.61)	3.56 (2.35)	–
Single-choice test	6.97 (2.26)	9.06 (1.22)	7.50 (2.74)	–	9.66 (0.70)

measurement was significantly more influential than observation, with the test parameter t value $t(31) = 2.08$ and p value $p = 0.046$. However, with Cohen's $d = 0.37$, this effect was small, according to [38].

The manageability was assessed for the measure carried out before and after the learning process. The results are presented in the last column of Table 6. These values primarily refer to the operation of the tablet.

In the open-ended part of the interview, 30 out of 32 subjects believed it is generally helpful to measure learning outcome. Motivation was frequently cited because measuring learning progress allows one's actions to be reflected and optimized. Furthermore, such measurements can be used for planning the further development of employees and ensuring high-quality and error-free results in assembly.

4.6 Discussion

The study carried out is limited by the fact that subjects were not employed in assembly. However, an attempt was made to represent as diverse a group as possible in line with the diverse employee structure in manual assembly, as mentioned in Sect. 1. This was done by selecting people of different ages and with different levels of experience. In this context, the level of education differs from that of the group of assembly employees, as many students participated in the study. Therefore, the qualitative conclusions drawn from the interview in the study need to be critically reviewed. The aim of the study, to prove the quality of the evaluation measures, can be seen independently of the group, so that the results can be used further.

All the evaluation measures show a statistically significant increase in competence during the learning process. It was shown that the tests alone are meaningful. However, several tests together had a higher significance. This is in line with the definition of competence development, in which competencies are described as the sum of different action-oriented skills [6]. The different competencies can be evaluated with different measures. In conclusion, it is important to select appropriate tests for learning situations according to the selection process in Sect. 4.

5 Conclusion

This paper presents a procedure for selecting learning outcome evaluation measures, the implementation of various measures in the learning factory for the study, and the practical application by means of an empirical study in an assembly setting. The study shows that learning among assembly workers can be effectively measured using measures as competence self-assessment, time measurement, observation, and single-choice test.

Moreover, the findings highlight the importance of tailoring evaluation strategies to meet the dynamic needs of assembly workers. The implications suggest potential for more extensive studies involving assembly personnel, to gain a deeper understanding of their learning processes and preferences. Such studies are fundamental to the development and implementation of optimized learning programs. Additionally, this paper advocates future research efforts aimed at streamlining the design and implementation of learning processes. Given the fast-paced and often high-pressure production environments, such as those caused by mass customization, that characterize assembly work, the need for simplified and faster methods is emphasized.

In conclusion, this study contributes to assessing learning among assembly workers by providing insight into effective evaluation measures while indicating a clear direction for future research to refine and optimize learning processes in industrial settings.

Acknowledgements The authors thank the German Federal Ministry for Economic Affairs and Climate Action (BMWK) for its financial and organizational support of the 'Mittelstand-Digital Zentrum Augsburg' [Grant no. 01MF22002B].

Author contributions M.M. set up the idea and research objectives. M.M. and K.S. conducted the literature review and the study, and wrote the main text of the manuscript. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest No potential conflict of interest was reported by the authors.

References

1. Breque M, Nul L de, Petridis A (2021) Industry 5.0: Towards a sustainable, human-centric and resilient European industry. R&I Paper Series, policy brief. Publications Office of the European Union, Luxembourg
2. Fink K (2020) Cognitive Assistance Systems for Manual Assembly throughout the German Manufacturing Industry. J Appl Leadership Manag 8:38–53
3. German Chamber of Industry and Commerce (2023) Uncertain framework conditions slow down German economy: DIHK Economic Survey Fall 2023. https://www.dihk.de/resource/blob/105390/856886caef029e94809dcf720335bede/economic-survey-fall-2023-data.pdf. Accessed 23 Nov 2023
4. Rudy B (2022) Build Learning into Your Employees' Workflow. https://hbr.org/2022/07/build-learning-into-your-employees-workflow. Accessed 25 Oct 2023
5. Burggräf P, Dannapfel M, Adlon T et al (2021) Adaptive assembly systems for enabling agile assembly—empirical analysis focusing on cognitive worker assistance. Procedia CIRP 97:319–324. https://doi.org/10.1016/j.procir.2020.05.244
6. Maier M, Schulz J (2023) Concept for the competence development and learning process of assembly workers. In: 2023 IEEE international conference on industrial engineering and engineering management (IEEM). IEEE
7. Cedefop (2023) European guidelines for validating non-formal and informal learning. Cedefop reference series, vol 124, 3rd edn. Publications Office of the European Union, Luxembourg
8. Sehr P, Moriz N, Heinz-Jakobs M et al (2022) Am i done learning?—determining learning states in adaptive assembly systems. In: 2022 IEEE 27th international conference on emerging technologies and factory automation (ETFA). IEEE, pp 1–8
9. North K, Kumta G (2014) Knowledge Management. Springer International Publishing, Cham
10. Euteneier RJ, Scheelen FM (2017) ASSESS by SCHEELEN, ASSESS performance analyse, ASSESS kompetenzanalyse, ASSESS 360-grad-analyse—entwicklung von kompetenzmodellen, management development, strategisches kompetenzmanagement. In: Erpenbeck J, von Rosenstiel L, Grote S et al (eds) Handbuch Kompetenzmessung: erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 3rd edn. Schäffer-Poeschel, Stuttgart, pp 136–148
11. Gay F, Wittmann R (2017) Persolog persönlichkeits-profil—schlüssel-potenziale für kompetenzen. In: Erpenbeck J, von Rosenstiel L, Grote S et al (eds) Handbuch kompetenzmessung: erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 3rd edn. Schäffer-Poeschel, Stuttgart, pp 174–189
12. Heyse V (2017) KODE und KODEX - Kompetenzen erkennen, um Kompetenzen zu entwickeln und zu bestärken. In: Erpenbeck J, von Rosenstiel L, Grote S et al (eds) Handbuch Kompetenzmessung: Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 3rd edn. Schäffer-Poeschel, Stuttgart, pp 245–273
13. Kaufhold M, Barthel C (2017) LERNSTÜCK - Ein Verfahren zur Zertifizierung informell erworbener Kompetenzen auf Basis dokumentierter Arbeitsprozesse. In: Erpenbeck J, von Rosenstiel L, Grote S et al (eds) Handbuch Kompetenzmessung: Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 3rd edn. Schäffer-Poeschel, Stuttgart, pp 346–354
14. Mollet A (2017) COMPRO+ Competence Profiling. In: Erpenbeck J, von Rosenstiel L, Grote S et al (eds) Handbuch Kompetenzmessung: Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 3rd edn. Schäffer-Poeschel, Stuttgart, pp 430–440
15. Schulz R, Schardien P, Hossiep R (2017) Das bochumer inventar zur berufsbezogenen persönlichkeitsbeschreibung (BIP). In: Erpenbeck J, von Rosenstiel L, Grote S et al (eds) Handbuch Kompetenzmessung: erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 3rd edn. Schäffer-Poeschel, Stuttgart, pp 580–596
16. Hertle C, Tisch M, Kläs H et al (2016) Recording shop floor management competencies—a guideline for a systematic competency gap analysis. Procedia CIRP 57:625–630. https://doi.org/10.1016/j.procir.2016.11.108
17. Hambach J, Tenberg R, Metternich J (2015) Guideline-based video analysis of competencies for a target-oriented continuous improvement process. Procedia CIRP 32:25–30. https://doi.org/10.1016/j.procir.2015.02.212
18. North K, Reinhardt K, Sieber-Suter B (2018) Kompetenzmanagement in der Praxis: Mitarbeiterkompetenzen systematisch identifizieren, nutzen und entwickeln : mit vielen Praxisbeispielen, 3rd edn. Springer Gabler, Wiesbaden
19. Glass R, Metternich J (2020) Method to measure competencies—a concept for development, design and validation. Procedia Manuf 45:37–42. https://doi.org/10.1016/j.promfg.2020.04.056
20. Arena D, Perini S, Taisch M et al (2018) The training data evaluation tool: towards a unified ontology-based solution for industrial training evaluation. Procedia Manuf 23:219–224. https://doi.org/10.1016/j.promfg.2018.04.020
21. Perini S, Arena D, Kiritsis D et al (2017) An ontology-based model for training evaluation and skill classification in an industry 4.0 environment. In: Lödding H, Riedel R, Thoben K-D et al (eds) Advances in production management systems. The path to intelligent, collaborative and sustainable manufacturing, vol 513. Springer, Cham, pp 314–321
22. Kuna P, Hašková A, Hodál P (2022) Tailor-made training for industrial sector employees. Sustainability 14:2104. https://doi.org/10.3390/su14042104
23. Gross E, Siegert J, Tenberg R et al (2022) Extension of assembly system planning methods to include competence development in the value-added process. SSRN J. https://doi.org/10.2139/ssrn.4071985
24. Wilschut ES, Könemann R, Murphy MS et al (2019) Evaluating learning approaches for product assembly. In: Makedon F (ed) PETRA 2019: the 12th ACM international conference on pervasive technologies related to assistive environments. ACM, New York, pp 376–381
25. Blickle G (2019) Leistungsbeurteilung. In: Nerdinger FW, Blickle G, Schaper N et al (eds) Arbeits- und organisationspsychologie, 4th edn. Springer, Berlin, pp 303–323
26. Landy FJ, Farr JL (1980) Performance rating. Psychol Bull 87:72–107. https://doi.org/10.1037/0033-2909.87.1.72

27. Ferris GR, Witt LA, Hochwarter WA (2001) Interaction of social skill and general mental ability on job performance and salary. J Appl Psychol 86:1075–1082. https://doi.org/10.1037/0021-9010.86.6.1075
28. Morgan SJ, Pullon SRH, Macdonald LM et al (2017) Case study observational research: a framework for conducting case study research where observation data are the focus. Qual Health Res 27:1060–1068. https://doi.org/10.1177/1049732316649160
29. Werrlich S, Lorber C, Nguyen P-A et al (2018) Assembly training: Comparing the effects of head-mounted displays and face-to-face training. In: Chen JYC, Fragomeni G (eds) Virtual, augmented and mixed reality, 10909 LNCS. Springer, Cham, pp 462–476
30. Meyerhoff J (2016) Fachwissen lebendig vermitteln: Das Methodenhandbuch für Trainer und Dozenten, 4th edn. Springer, Wiesbaden
31. Bontis N, Hardie T, Serenko A (2009) Techniques for assessing skills and knowledge in a business strategy classroom. IJTCS 2:162. https://doi.org/10.1504/IJTCS.2009.031060
32. Yardley-Matwiejczuk KM (1997) Role play: theory and practice. Sage Publications, London
33. Riffenburgh R, Gillen D (2020) Statistics in Medicine, 4th edn. Elsevier, Oxford
34. Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Human mental workload, vol 52. Elsevier, Oxford, pp 139–183
35. Brooke J (1995) SUS: A quick and dirty usability scale. Usabil Eval Ind 189:4–7
36. Jeske T, Schlick CM (2013) a new method for forecasting the learning time of sensorimotor tasks. In: Trzcielìnski S, Karwowski W (eds) Advances in ergonomics in manufacturing. CRC, Boca Raton, pp 241–250
37. Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the system usability scale. Int J Hum Comput Interact 24:574–594. https://doi.org/10.1080/10447310802205776
38. Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Mahwah