Empirical Research in
Vocational Education and Training

RESEARCH Open Access

Digital measurement of hands-on performance? Ecological validation of a computer-based assessment of automotive repair skills

Stefan Hartmann[1*], Emre Güzel[1] and Tobias Gschwendtner[1]

*Correspondence:
stefan.hartmann@ph-ludwigsburg.de

[1] Abteilung Technik und ihre Didaktik, Pädagogische Hochschule Ludwigsburg, Reuteallee 46, 71634 Ludwigsburg, Germany

Abstract

We investigated the ecological validity of performance measures from a computer-based assessment tool that utilises scripted video vignettes. The intended purpose of this tool is to assess the maintenance and repair skills of automotive technician apprentices, complementing traditional hands-on assessment formats from the German journeymen's exams. We hypothesise that the ability to correctly judge repair actions shown in videos is a good predictor of the ability to perform corresponding actions in hands-on scenarios. Apprentices in the third year of vocational training carried out repairs on real cars or car systems, while experts rated their performance. After this, they worked on our computer-based tests, which utilise videos of very similar repairs. The correlation between video judgement and hands-on performance was lower than expected for most repair actions as well as for overall scores, indicating insufficient ecological validity of the test score interpretations. However, the findings are promising for developing future tests, as the results for some repair actions indicate it is generally possible to develop ecologically valid video-based items focusing on hands-on skills. We discuss the results in the light of a validation framework that combines validity evidence from different sources for the same assessment tool. Finally, we hope our findings contribute to a broader discussion about the psychometric quality of exams.

Keywords: Ecological validity, Validation, Authentic assessment, Hands-on skills, VET

Introduction

In vocational education and training (VET), computer-based assessment (CBA) environments enable teachers, trainers, and organisations to economically test job-related skills and abilities of individuals and groups of learners (Conole and Warburton 2005; The International Test Commission [ITC], 2006; Malone 2020). CBAs can be used in classrooms, at companies, or at home to monitor learners' skill levels as well as their learning progress; they can also complement or even replace traditional exam formats during or at the end of apprenticeship or other vocational training (The Commission

on Technology and Adult Learning [CTAL] 2001). Especially when aiming at learners' practical expertise and hands-on skills, CBAs are a promising alternative to direct observations of job performance. Standardised CBAs have the potential to combine technological advantages such as audio-visual aids, interactive elements, or adaptive item selection with automated scoring, unmatched objectivity, and high reliability of the test results (Kirschner et al. 2017; Williamson et al. 2006). One particular benefit that makes CBAs attractive is implementing innovative item types that exceed the often limited possibilities of traditional assessment formats (Parshall et al. 2010). By incorporating realistic tasks, CBAs can be most authentic (i.e., ecologically valid), while a high degree of standardisation ensures good reliability and internal validity. These criteria are much more difficult to achieve with observational studies (Kirschner et al. 2017).

Validity and authenticity

Authentic assessment items often emulate situations similar to those of the real world and therefore engage students or apprentices in cognitive processes "under the same working conditions … as they would have had in life beyond school" (Palm 2008, p. 6). However, authenticity has its limits. Even the most innovative, most sophisticated test item is only an approximation of reality—a model that provides test users with information that allow them to extrapolate judgements about skills which are difficult, expensive, dangerous, or even impossible to assess otherwise. To fit the intended purpose of a CBA—or any other assessment for that matter—its authors should provide evidence their interpretation of the test results is both adequate and appropriate. The process of collecting and discussing such evidence is known as *validation* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 2014; Rupp and Pant 2006).

Messick (1987) describes validation as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (p. 1, emphasis in the original). There is consensus that any claim of validity needs to be justified by appropriate evidence from various sources (Kane 2016). According to the *Standards of Educational and Psychological Testing* (AERA, APA and NCME, 2014), such sources can be test content, response processes, internal structure, relationships with conceptually related constructs, relationships with criteria, and consequences of the test. While there are different approaches and theoretical frameworks regarding the details of validity and validation, there is broad agreement that at least *some* validity evidence must be presented for any educational and psychometric assessment—not just in scientific research, but also in practise where assessment results are often used to determine school admission, final degrees, or job qualifications, and therefore directly affect people's lives and professional careers.

Several authors, however, have found that in practice, validation is often neglected or done inadequately. Gafni (2016) criticises that numerous tests used for selection or admission purposes lack sufficient validity evidence. Cook et al. (2013) reviewed 417 technology-enhanced instruments and concluded that validity evidence "is sparse [and] leaves much room for improvement" (p. 872). As of 2023, the situation has not significantly improved. Researchers in the educational sciences still rarely report any

validity evidence at all, or fail to make the important distinction between a test, test scores, and test score interpretations. In the Book of Abstracts of the 2023 EARLI conference, Europe's biggest professional meeting of educational scientists, the vast majority of validation studies refer to a validation of "scales", "tests", "instruments", "assessments", "data", "measures" etc. (European Association for Research in Learning and Instruction [EARLI], 2023). In only three out of 42 studies in which validation was carried out, the validity of test score *interpretations* was investigated. In most of the studies, evidence from a single source was used to investigate validity.

The importance of validation especially applies to CBAs, as international test standards and guidelines stress "test users should not rely solely on computer-generated interpretations of test results" (AERA, APA and NCME 2014, p. 144) and "advanced multimedia features should be used only where justified by validity" (ITC 2006, p. 147).

When assessments are designed to incorporate authentic tasks, one particularly important aspect of validity is *ecological validity*. Ecological validity refers to "the degree to which test performance predicts behaviors in real-world settings" (Gouvier et al. 2010, p. 399). It therefore depends on how similar the context of the assessment tasks is to the real-life tasks the assessment is aiming at. Even though often overlooked or entirely neglected, ecological validity is an essential prerequisite for raising the acceptance of psychometrically sound test instruments amongst job professionals, educators, and examiners in the vocational education, training, and examination system. Such practitioners sometimes tend to distrust evidence from mere experimental and/or lab settings that often appear artificial and constructed when compared to real-world situations (Kingstone et al. 2008). With CBA, a testing mode effect is added to these concerns, and some argue that comparability does not hold "when examinees are tested in a mode different from the one in which they routinely work" (Bennett 2002, p. 13). The criticism boils down to questions such as:

- To what degree can clicks on a computer reflect a person's ability to troubleshoot a malfunctioning car engine?
- How well do multiple-choice items measure practical skills like maintaining a hydraulic brake?
- Is it even possible to validly computer-simulate hands-on tasks like fixing a torn electrical wire with a solder connector?

Though these questions are intuitively reasonable and justified, the concept of ecological validity has not been undisputed, and its practical applications have been criticised for lacking specificity or falling short of addressing the problem of generalisability (Holleman et al. 2020). Furthermore, ecological validity is neither mentioned explicitly in the *Standards of Educational and Psychological Testing* (AERA, APA and NCME 2014) nor in the *International Guidelines on Computer-Based and Internet-Delivered Testing* (ITC 2006). However, it can be subsumed under predictive validity, which is part of criterion-related validity (Stieler 2011). Aiming at the question of how well assessment results predict performance in operationally distinct real-world situations, ecological validity fits the definition of test-criterion relationships as a source

of validity evidence that can "be evaluated in terms of the accuracy with which a test score could predict or estimate the value of … an observable performance measure" (Rupp and Pant 2006, p. 1033).

Designing authentic assessment items using video vignettes

In the study presented in this paper, we evaluate the accuracy with which the test scores from a CBA predict a hands-on performance measure of automotive maintenance and repair skills. The CBA does not require test takers to carry out repairs, but rather to judge repairs presented in scripted video vignettes on a screen. Interpreting the test takers' judgements of the videos as indicators of their hands-on skills means inferring from a measure of procedural knowledge about a task to the ability of performing this task in an operationally distinct real-world situation. This raises the question of ecological validity.

The target population of the CBA are automotive technicians in Germany. Automotive technicians ("Kfz-Mechatroniker*innen" in German) service and maintain passenger cars and light service vehicles. They carry out routine service tasks as well as case-specific diagnoses and repairs. The skills required to successfully perform their job "range from replacing simple parts to solving complex faults using diagnostics equipment" (The Transport Training Board 2023). With automotive technology rapidly evolving over the past decades, skills in electrical engineering, information technology, and high-voltage technology were added to the job profile of the traditional "hands-on" mechanic.

In Germany, where apprenticeship programmes are a combination of workplace learning and classroom-based learning (Rausch et al. 2016), apprentice automotive technicians must pass an exam ("Gesellenprüfung" in German) which is split into a theoretical (written) part and a practical (hands-on) part. The hands-on part consists of several stations where the apprentices must demonstrate their maintenance and repair skills to an expert judge. From a psychometric perspective, these stations resemble a performance assessment (cf. AREA, APA, NCME 2014, p. 77).

The rating method of one observer rating one examinee could be described as *human scoring* (Bejar et al. 2006). Due to their low level of standardisation and high logistical requirements, performance assessments using human scoring are very expensive and at the same time often associated with poor objectivity and reliability (Bejar et al. 2006; Stecher and Klein 1997; Weber et al. 2015). At each exam station, only one examiner rates the examinees' performance; therefore, evaluating objectivity based on an empirical measure such as inter-rater agreement is impossible.

The COVID-19 pandemic increased the difficulties of using human examiners at the exams. Due to changed regulations, for example, regarding social distancing and the examination room requirements, but also as a consequence of rising absenteeism due to sick leave, the demand for examiners has increased even further (Deutscher Industrie- und Handelskammertag [DIHK], 2021). As result of these issues, the need for a more standardised, less expensive, and at least to some degree automated alternative to human scoring in the exams is constantly increasing. In reaction to this demand, we developed digital assessment tools to examine automotive technician apprentices' hands-on performance during their final exams. The first tool is a computer simulation, aiming at diagnostic skills with a focus on electric and electronic car systems (Gschwendtner et al.

2009; Norwig, et al. 2021). The second tool is a video-vignette based test, designed to assess skills that are more closely related to repair tasks that require "traditional" hands-on work (Gschwendtner et al. 2017). The test uses videos of authentic and holistic tasks derived from core work processes in this domain (The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany [KMK], 2009; Spöttl et al. 2011).

Before these tools can be used in exams, it is imperative to investigate the validity of the test score interpretations that are based on the results. In accordance with the *Standards for Educational and Psychological Testing*, we conducted a series of validation studies, collecting and discussing evidence from various sources:

- The representativeness of both CBA instruments was ensured by a close involvement of professional experts to the item and test construction process. These experts are master craftsmen and journeymen, automotive engineers, vocational school teachers, and experienced examiners.
- Test content for both tools is based on the content of real-world job tasks as well as journeymen's exam stations.
- Answering options for the single-best-answer item format of our video-based assessment were constructed based on the written responses given by automotive technician apprentices during a previous study, ensuring that both the correct options, as well as the distractors, realistically and plausibly represent common mental concepts of the target group (see Sadler 1998, for a comprehensive description and discussion of such a distractor-driven item construction approach).
- The authenticity and appropriateness of the scripted video vignettes and related single-best-answer items were rated by experts.
- For our simulation-based assessment instrument, we carried out cognitive labs to explore the thinking processes during item solving (Norwig et al. 2021).
- For our video-based tool, we conducted a known-groups validation study (Hartmann and Gschwendtner 2021).
- The ecological validity of our simulation-based tool was investigated in multiple studies (Gschwendtner et al. 2017).

However, none of these measures and studies aimed at the ecological validity of our video-based assessment tool. Before the tool can be used in exams, the following research question should be answered: To what degree do measures of repair skills based on scripted video-vignettes validly predict hands-on performance?

Method

Video-based assessment

The aim of this study is to investigate the ecological validity of the CBA that uses video vignettes to assess repair skills of automotive technician apprentices. The instrument is a computerised fixed test (CFT), meaning that the same test items are provided to all participants in a fixed order without adaptive item selection (Parshall et al. 2002). Its intended purpose is to be used in the practical part of the final exams for automotive technicians in the German VET system, possibly complementing or replacing existing

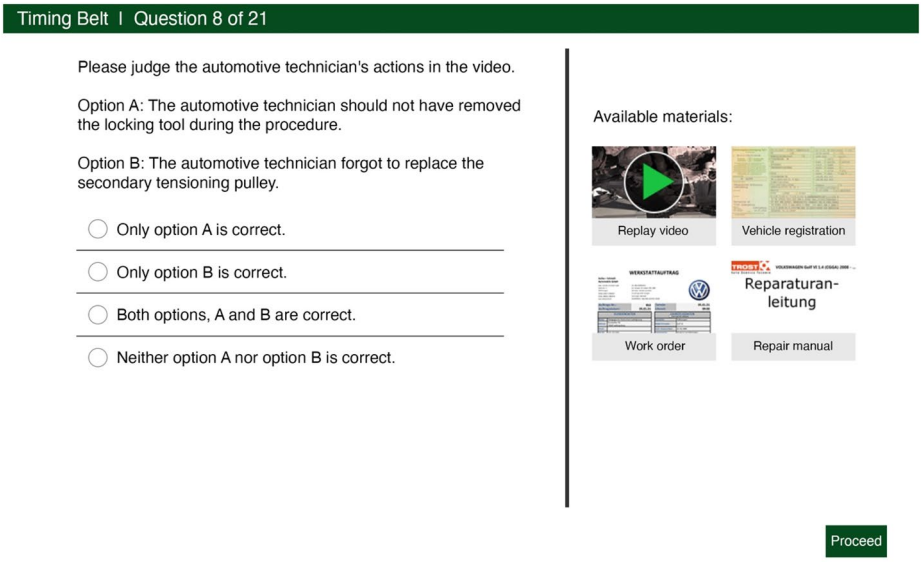

Fig. 1 One of the items used in the video-based CBA (left). After an automatic playback of the video has finished, the video can be played again by clicking the thumbnail at the right side of the screen. Additional materials such as excerpts from a repair manual are also available in the right screen section

hands-on exam stations. We concentrated our efforts on an item design that combines video vignettes with questions in a single-best-answer format. Such an item design, known as *scripted video vignettes*, has been successfully used in the past, e.g., to assess teachers' subject-specific competences (Knievel et al. 2015), to measure commercial knowledge and skills (Rohr-Mentele and Forster-Heinzer 2021), or to investigate patient-provider communication in the health sector (Hillen et al. 2013).

Scripted video vignettes are "short visual depictions of pre-written (hypothetical) events" (Hillen et al. 2013, p. 296). In our study, the scripted video vignettes are short film sequences in which an actor plays an automotive technician who performs different repair tasks in a realistic manner. We filmed videos that depict repairs on five different automotive systems: electric lighting, hydraulic brakes, exhaust and catalytic converter, suspension and steering, and engine timing belt. The video vignettes for the five systems are used as a basis for five independent test blocks.

The video vignettes are between 6 s and 3:10 min in length; most of the clips are less than one minute long. Each of the vignettes shows a single repair step at a time. Each repair step consists of one or several actions, such as tightening a bolt, measuring voltage with a multimeter, or applying a lubricant to a surface. Some of the actions shown in the videos are performed correctly by the actor, while others intentionally depict errors and mistakes either known to be common for automotive technician apprentices or strongly assumed by experts who have a long record of working with such apprentices (Fig. 1).

Each video is followed by one or several text-based selected-response items displayed on the computer screen. We used a highly standardised "all possible options" (APO) item design (Moon et al. 2019). In every item, two short text statements—labelled *Option A* and *Option B*—are presented. Both statements refer to the repair actions shown in the video, e.g., "The technician in the video should have replaced the brake hose" or "The

multimeter in the video was set to the wrong scale range." These statements were generated using response data from a previous study in which apprentice automotive technicians watched the same videos and judged the technician's actions in the form of written responses. We revised and categorised their responses, then identified the most common answers, assuming these answers represent widespread concepts about the tasks shown in the videos. The answers, carefully revised in terms of length, spelling, and grammar, were then presented as options A and B to choose from. In some items, one or both options are *correct* statements, representing conceptions that reflect the knowledge and skills already acquired by the target group. One or both options can also be *incorrect* statements, reflecting common misconceptions. To rate whether the options are correct or incorrect, the participants must select one of four possibilities:

1. Only option A is correct,
2. Only option B is correct,
3. Both options are correct, or
4. Neither option is correct.

Thus, all test items consist of the same four answering categories to choose from, with only one correct answer (1, 2, 3, or 4).

In comparison to traditional multiple-choice items with four true or false statements, this format is expected to reduce reading effort and cognitive load. Distractors that represent common misconceptions are more plausible to test takers than artificially constructed false options. The APO format also affects test-takers' response tendencies under uncertainty. By explicitly presenting the two options "both" and "neither", test takers' false assumptions that one of the two statements must always be correct and the other incorrect is significantly reduced (Moon et al. 2019).

Assuming that one can only differentiate between correct and incorrect repair actions if one knows how these repairs are adequately performed, the total number of correct responses can be interpreted as an indicator of repair-related procedural knowledge.[1] To judge the ecological validity of this test score interpretation, we compared the results of the video-based CBA to hands-on performance measures from stations that required apprentices to carry out repairs highly similar to those in the videos. Three of the vignette-based test blocks were used for this validation study: electric lighting, hydraulic brakes, and timing belt (Fig. 2).

Hands-on performance assessment

To investigate how well the judgement of repair tasks in our scripted video-vignettes predicts the apprentices' skill of performing these tasks in an actual, real-world scenario, we designed a performance assessment using real cars (or car systems, respectively). We acquired three cars of the same model used in our video vignettes, and also two engines of the same type. Furthermore, we designed two identical models of the car's front lighting system, using real headlights of the same type shown in our

[1] This procedural knowledge should be closely related to the ability of practically performing these tasks; however, it is possible that someone knows how to do something without being able to do it by themself.

Fig. 2 Layout of the repair stations used in the validation study. (Image credit: Emre Güzel | https://creativecommons.org/licenses/by-sa/4.0/)

videos. We manipulated the cars, car engines, and lighting models with mechanical and electrical faults that were very similar (though not perfectly identical) to the faults the actor in our scripted videos attempted to repair.

For each station, we prepared a list of several repair requests, such as "Fix the broken electrical wire", "Replace the brake pads", or "Disassemble the timing belt from the engine". Examiners were assigned to the stations to observe the participants' repair actions and rate them using a highly standardised protocol. Most examiners were master craftsmen with a record of serving as volunteers in the annual journeymen's exams. Additional examiners were scientists familiar with the project. To ensure that all examiners rated the examinees in the exact same manner, they were instructed in a training workshop prior to the study.

The standardised protocols used in the study contained an extensive list of possible actions related to each repair task. Some of these actions were considered appropriate (correct) in the context of the task, and some inappropriate (incorrect). The examiners were instructed to mark each action in checkboxes labelled with "Yes" (i.e., action was performed by participant) and "No" (action was not performed). By simply documenting which actions they observed—instead of *judging* the actions—the margin for interpretation by the examiners was reduced to a minimum.

For each repair task, there was also a free-text field on the protocol in which the examiners could note additional observations. Each protocol consisted of four to five short repair tasks, the list of applicable actions, and free-text fields. On the last page of each protocol, the examiners were asked to rate the participant's overall

performance, competence, and level of routine using Likert-type items, and finally to judge the observed repair job in terms of a school grade.

In deviation from the CBA, a generous time limit of 35 min (including instructions) was set for each station. The limit was necessary to ensure a seamless transition between the exam stations for all participants and to avoid long waiting queues. The given time was more than sufficient to complete the repair tasks at a reasonable work pace. However, several participants did not finish in time at the *timing belt* station. This was to our surprise, as we had discussed that station with experienced examiners who informed us that a very similar station in the practical exams consists of two tasks instead of just one and is frequently solved within 30 min. According to the protocols, apprentices who struggled to finish in time on this station showed significant signs of uncertainty and reported that they were unfamiliar with the task they were asked to perform. Overall, the design of the stations used in this performance assessment was very similar to that of the stations used during the practical part of the final exams for automotive technicians in Germany.

Sample and procedure

The assessment took place November 2021 in the training facilities of the Chamber of Crafts of the Stuttgart region in southwestern Germany. Participants were determined by approaching vocational schools in the region. We invited entire third-year courses of automotive technician apprentices to participate in our study. Participation was voluntary both at the school level as well as the individual level. All potential participants were informed about the scientific nature and overall purpose of the study, the entirely anonymous character of data collection, and the intended use of the collected data. It was stressed that none of the data would be made available to third parties outside the scientific community, and the decision not to participate in the study or non-attendance would not result in any disadvantages.

To keep the effort for the participants at a reasonable level, each participant had to solve hands-on tasks at only two stations. Using a balanced incomplete block design, we assigned the same number of apprentices to each possible combination and timely order of two hands-on stations (lighting + brakes, brakes + lighting, timing belt + brakes, brakes + timing belt, timing belt + lighting, lighting + timing belt).

Each participant was summoned to attend at a pre-arranged time. On arrival, they were handed a personal schedule with an anonymous ID code, containing the names of the stations they had been assigned to as well as the times when to attend those stations. After they had completed the tasks at their two hands-on stations, the participants were sent to a computer pool where they took the vignette-based tests *for the same two car systems* they just had worked on. We decided to conduct the assessment in this successive order to prevent the participants' hands-on performance being affected by actions from the videos. Because the apprentices were given no feedback on their performance at the hands-on stations, the risk of memory effects affecting their response tendencies in the CBA was minimal.

Each hands-on station and the computer pool were equipped with a station schedule to ensure that each participant was at the right place at the right time. Our balanced incomplete block design allowed for a maximum number of 120 apprentices; however,

Table 1 Station overview and sample sizes

Car system	Number of hands-on repair tasks	Number of potential repair actions listed on the protocol	Number of video-based items	n
Electric lighting	4	23	17	47
Hydraulic brakes	5	42	18	39
Timing belt	4	42	17	40

not all available time slots were filled. Therefore the number of participants per station varied from 39 to 47 apprentices (see Table 1 for a complete overview).

Statistical analysis

For each item in the video-based CBA, correct responses were credited with a score of 1 point (full credit) whereas incorrect responses were scored 0 points (no credit). The total number of correct responses reflects a test taker's ability to correctly judge the actions shown in the videos, representing their procedural knowledge regarding the repair tasks on that car system.

To rate the examinees' hands-on performance, each repair action necessary and appropriate to complete the corresponding repair task marked with a "Yes" on the protocol (i.e., action performed) was credited with 1 point (full credit), whereas necessary and appropriate actions marked with a "No" were scored 0 points (no credit). In line with the scoring method used in the CBA as well as in real exams, we did not use penalties (i.e., negative scores) for incorrect, inappropriate, or unnecessary actions. The total number of necessary and appropriate actions performed by the examinees reflects their ability to correctly carry out the repair tasks at this station.

The rating system of both instruments allows comparing the participants' ability to judge repair actions in the videos with their ability to perform comparable actions in the hands-on assessment. The comparison is carried out at the item level as well as the level of total scores for each of the three car systems (lighting, brakes, timing belt).

To investigate to what degree the apprentices' judgements of repair actions in the video-based assessments predict their performance of comparable actions at the hands-on stations, we carried out an item-wise analysis first. For each item in the video-based test, we identified the repair action in the video vignette that a certain statement—presented as Option A or Option B in the item—refers to. We recoded the item scores (answers 1, 2, 3, or 4), so that a score of 1 is associated with either the *agreement* to a statement that reflects an *appropriate* action performed by the actor in the video, or with *disagreement* to a statement that reflects an *inappropriate* action. This score can be compared directly to the examiner's markings on the protocols for the corresponding repair action at the hands-on stations. It is hypothesised that correct judgements of the videos correspond with correct hands-on performance, and incorrect judgements correspond with incorrect performance (see Fig. 3).

To investigate how well the overall scores from each of the three video-based tests predict the overall performance at the three corresponding hands-on stations, we carried out correlational analyses.

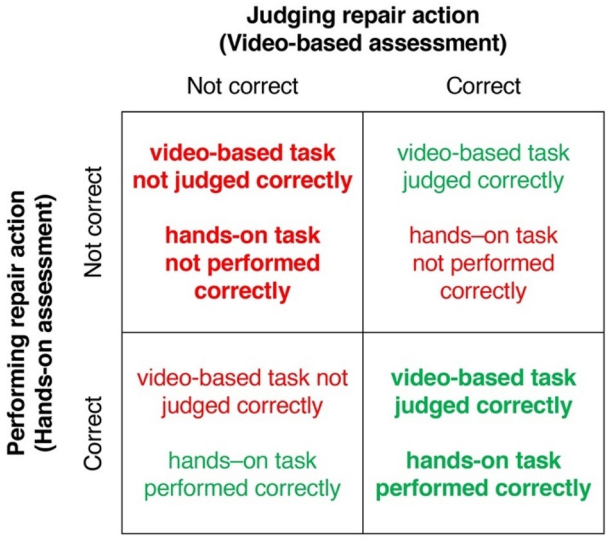

Fig. 3 It is hypothesised that the judgement of repair actions shown in video vignettes predicts the outcome of performing comparable actions in a hands-on assessment. The upper left cell and lower right cell of the contingency table—printed in bold—mark the hypothesised outcome. The upper right cell and lower left cell mark situations in which the results of the video-based test do not predict hands-on performance

All statistical analyses were carried out in R version 4.0.2 (R Core Team 2020), mostly using the standard *stats* package. Package *psych* version 2.0.12 (Revelle 2020) was used for descriptive analyses. Charts were created using the packages *gplot* version 3.1.1 (Warnes et al. 2022) and *ggplot2* version 3.3.6 (Wickham 2016).

Results

The CBA test blocks consist of very heterogeneous item pools. Accordingly, the discrimination parameters of most of the video-based items were low; some were close to zero or even had negative values. Cronbach's α ranges from 0.23 to 0.50, and Guttman's λ-6 (which is more appropriate for tests in which the association to the underlying variable is varying across the items) ranges from 0.39 to 0.61.[2]

Predictive potential of video-based items

Not all repair actions shown in the video vignettes were part of the tasks that had to be performed in the hands-on assessment, and not all actions necessary to successfully carry out a repair were depicted in the video vignettes. Therefore, the association between video judgement and hands-on performance could only be evaluated at the item level for those repair actions from our video-based tests closely associated to a repair action in the hands-on assessment. We generated contingency tables, the general layout of which is explained in Fig. 3. The two cells marked with bold text indicate the hypothesised association, i.e., cases in which an apprentice's judgement

[2] Both parameters, α and λ-6, are indicators not of reliability per se, but of internal consistency, and it could be argued that this notion of reliability might be of limited use or even inappropriate for tests such as exams which cover a wide range of various abilities and therefore are not internally consistent by design. However, low internal consistency directly affects the outcome of analyses such as correlation tests, and thereby also the inferences based on such tests.

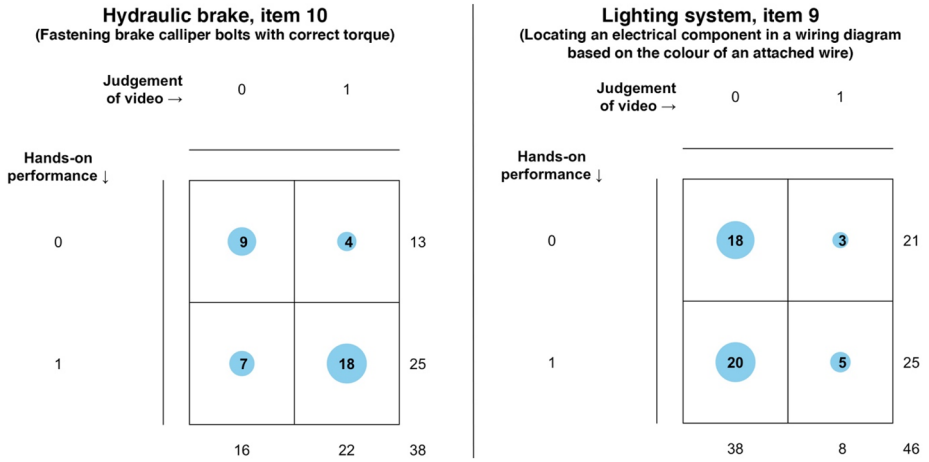

Fig. 4 For some repair actions, video judgements predict the outcome of the corresponding hands-on task to a certain degree (left; degree of association = 71.1%). For other repair actions, video judgement and hands-on performance do not correspond beyond chance (right; degree of association = 50.0%)

Table 2 Degree of association between the judgement of repair actions in the video-based assessment and performance at corresponding hands-on station

Car system	Number of corresponding repair actions	Degree of association > 50%	Degree of association > 70%	Lowest degree of association (%)	Highest degree of association (%)
Electric lighting	8	6	3	39.1	91.3
Hydraulic brakes	16	12	4	12.8	78.9
Timing belt	7	4	1	20.0	87.2

of an action they saw in the video corresponds with their own hands-on performance. The other two cells contain cases in which the responses of the two instruments are not corresponding, meaning that an action from a video was judged correctly, but the test taker failed to successfully perform the action at the hands-on station, or vice versa: a video was not correctly judged, but the corresponding task was successfully performed during the hands-on assessment.

Under a perfect random distribution in which all cases are evenly distributed to the four cells of the table, each cell would be filled with 25% of the cases. The cases in the upper left cell and the lower right cell together would make up 50%. If performance in the CBA is associated to hands-on performance beyond chance, the share of cases in these two cells would exceed 50%. If video judgement was a *perfect* predictor of hands-on performance, 100% of the cases would be gathered in the two cells. We consider a value above 70% an indicator of substantial association (see examples in Fig. 4).

The degree of association ranges from 12.8% to 91.3% (see Table 2). It is larger than 50% for the majority of repair actions; however, an association larger than 70% is found only for 8 of 31 repair actions (see Table 2).

Predictive potential of total scores

To analyse the predictive potential of each of the video-based instruments to the apprentices' practical skills, we calculated the degree of association between the total scores of each of the three video-based instruments and the scores of the corresponding hands-on stations. The relationship between these variables is of a predictive nature and therefore could be modelled as regression. However, the data do not meet several conditions for linear regression: The association of the variables is not linear (see Scatterplots in Fig. 5), and the residuals aren't normally distributed, which in combination with the relatively small sample sizes might cause unwanted bias of the regression estimators and significance tests. Therefore, we calculated non-parametric correlations that allow estimating the association between the variables under the given conditions in a more robust way.

For all three car systems, we found the total scores of the video-based tests and the performance measures of the corresponding hands-on stations were positively related. But the effects are small, and the coefficients are not significantly different from zero, except for the station "timing belt" (see Table 3).

Discussion

We investigated to which degree judgements of scripted video vignettes depicting automotive repairs can be used to predict automotive technician apprentices' ability to successfully perform similar repair tasks in a realistic scenario. The results of three computer-based tests using video vignettes were compared with the repair performance at three hands-on stations. It was hypothesised that correct judgements of repair videos reflect procedural knowledge of the tasks shown in these videos, and that this procedural knowledge would be a strong predictor of successful repairs in a real-world scenario.

Analyses were carried out on an item basis as well as total score basis. At the item level, the hypothesis was supported only for a very limited number of repair actions. A substantial degree of association was found for 8 of 31 item pairs (video vs. hands-on). At the level of the total scores per test, the correlations were small, and a significant effect was found only for one of the three car systems ("timing belt"). These findings indicate that the test scores from the three video-based assessments used in this study are not an ecologically valid predictor of the skills necessary to successfully carry out the repair tasks at the three corresponding hands-on stations. However, the results of the item-based analyses indicate that some of the video-based items were good predictors of hands-on-performance, matching the outcome of the practical tasks up to 91.3%.

As possible causes for the small overall effects and inconsistent results at the item level we discuss:

- The item content,
- Potential testing mode effects,
- Potentially confounding covariates,

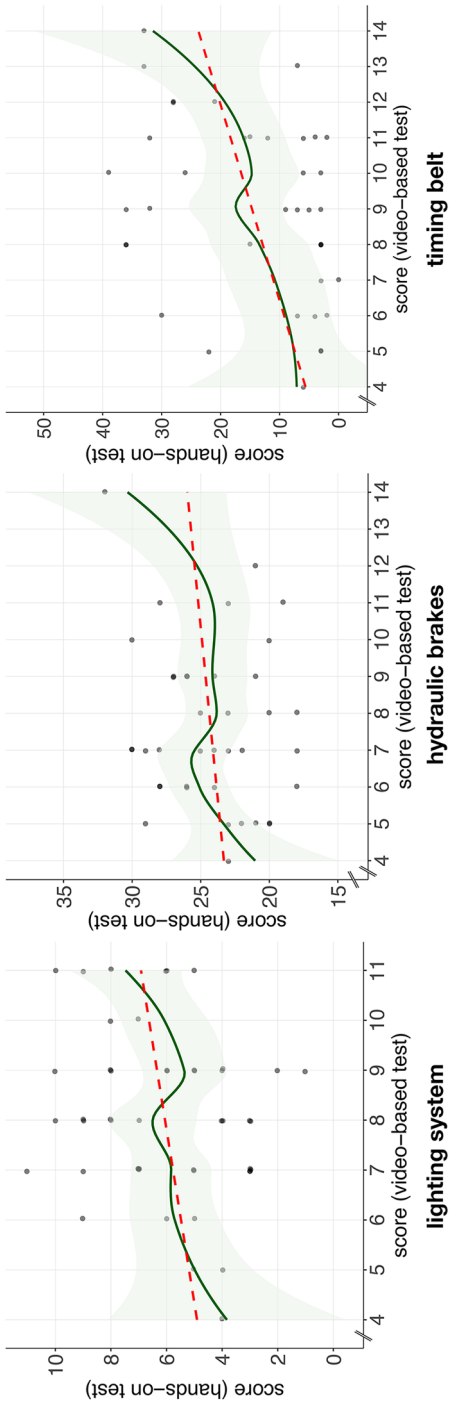

Fig. 5 Scatter plots indicate the relationship between the test scores of the video-based assessments (x-axes) and the hands-on assessments (y-axes) at the three stations "electric lighting", "hydraulic brakes", and "timing belt". The curved lines were fitted with the LOESS algorithm, whereas the dashed red lines indicate linear regression which seems less appropriate for these data

Table 3 Non-parametric correlations between the results of the video-based assessments and the performance at the corresponding hands-on stations

Car system	n	Kendall's τ	$p_{one\text{-}sided}$
Electric lighting	47	0.14	0.108
Hydraulic brakes	39	0.08	0.245
Timing belt	40	0.29	0.006

- Testing time,
- Memory effects,
- Reliability issues, and
- The validity of the hands-on tasks.

First, to lower the risk of learning effects between the video assessment and hands-on assessment, we did not re-create the repair tasks from our video vignettes in every detail for the hands-on stations. Instead, we developed hands-on tasks that were highly similar and, from an analytical point of view, most likely required the same set of skills to succeed. For example, the participants had to find an electric component in a circuit diagram based on the colours of the connecting wires. The same circuit diagram was used both in the video test and at the hands-on station, but different components and different wire colours were used for the tasks. It is possible—although unlikely—that such differences were crucial for success, therefore causing different outcomes in the video-based and hands-on assessments.

Another potential explanation is the *testing mode effect* (Bennett et al. 2008; Clariana & Wallace 2002) which states that even the same contents and materials used for assessment might produce different results when presented in different modes. Testing mode effects are sometimes (but not always and not consistently) present when the same test items are presented on paper vs. on a computer screen. In our study the video-based assessments were presented on computers and the hands-on assessments were presented as real-world repair tasks that had to be carried out on real cars or car systems. In the video-based assessment, basic computer skills were necessary to solve the items. At the hands-on stations, examinees had to work with real tools, adding haptic and motor skills to the tasks. It is therefore likely that a testing mode effect affects the way in which video judgements and hands-on repairs correlate. Furthermore, there is logical difference between judging an action and performing the same action. Procedural knowledge is a necessary condition to perform an action, but often not sufficient to do so.

Another difference between the two assessment formats emerges from the actions shown in the videos. Several actions performed by the actor show the correct way of carrying out a repair; therefore, during a video it might occur to apprentices that what they see is more adequate than the course of action they had previously chosen at the hands-on station. In such a situation, an incorrectly performed hands-on repair action would be accompanied by a correct answer to a video-based item.

The third explanation for the inconsistent findings and low correlations are potentially confounding variables. For example, reading abilities are known to play a significant role in assessments using text-based test items (Hartmann 2013). Our video-based assessment used text-based items after each video, requiring the participants to read and understand two short statements to judge the actions they just saw in the video. The repair tasks at the hands-on stations required little to no reading abilities, as the repair requests and other instructions were read out loud to the apprentices by the examiners. The only written materials they had to use were circuit diagrams and excerpts from repair manuals that contained information such as torque values. Again, however, the results do not support the assumption that reading abilities are accountable for the inconsistent findings.

At the 'timing belt' hands-on station, the limitation of testing time has potentially contributed to the low correlation of the total scores with those of the video-based test (which had no time limit). Several participants struggled with that station and did not finish the station in time. If the video-based test is used in a real exam, it should have the same time limit as the hands-on station.

Another potential reason for low correlations at the item level are memory effects. Even though there were no indications during the study, it is theoretically possible that failing to perform a certain task on a real car activates prior knowledge about that task. This prior knowledge might later help the examinee to find the correct answer to a corresponding item in the computer-based test.

The low correlations of the total scores and the inconsistent results at the item level are, of course, related to each other. If all video-based items were good predictors of the corresponding hands-on performance, the correlations at the scale level would be high as well. The large differences between the item results lead directly to the question of how well these items differentiate between individuals, and therefore, how reliable the scales built on those items are. The findings on internal consistency shed some light on this question. As one would expect for a test based on a highly heterogenous item pool, the item-total correlation parameters of the items and the internal consistency of the scales are not very high. To improve the internal consistency of the test, items with poor discrimination could be subsequently removed; however, this would in return significantly affect the test's ecological validity. If a video-based assessment holistically represents a typical car repair task, e.g., replacing the pads and disks of a car's brake, what would the test represent if the removal of the disk or the re-attachment of the brake caliper were removed from the test?

Finally, the ecological validity of the video-based assessment could be in perfect order, and the low and inconsistent association of the test scores to the outcomes of the repair tasks could be entirely the result of a low validity of the hands-on tasks used in this study. Indeed, little is known about the psychometric quality of hands-on exam stations. Though we used a highly standardised rating system, our system was still based on human examiners, and therefore is susceptible to individual decisions. Because the hands-on stations resembled real-world exam stations in many respects and suffered

from the same problems of finding volunteers to work as examiners, it was impossible to assign two examiners to each station and calculating inter-rater agreement to assess the objectivity of the protocols.

Without substantial knowledge about the psychometric properties of the test scores that were used as a criterion in this study, it is difficult to conclude which of the instruments is accountable for the low correlations: the video-based test, the hands-on stations, or both.

Relation to other validity evidence

It is noteworthy that validation is a process rather than the investigation of evidence from a single source (AERA, APA, NCME 2014). In our project, we honour that definition by continuing the collection of validity evidence from various sources, and by combining the findings to develop a comprehensive validation framework. It is part of this process to re-evaluate evidence from previous investigations of validity, and to set it into relation with the findings of this study. In a known-groups validation study carried out in 2021, we compared the test scores of automotive technician apprentices who processed our video-based instrument with the scores of apprentice electronics technicians (Hartmann and Gschwendtner 2021). The vehicle group outperformed the electronics group; significant medium to large group differences were found for all car systems investigated in the study. These findings indicate the skills required to correctly judge the repair tasks in the video vignettes are not of a general technical nature but related to the content domain of automotive engineering. The effect sizes varied between the tests (electric lighting: $d = 1.11$; hydraulic brake: $d = 0.61$; timing belt: $d = 1.07$). These differences somewhat resemble the differences found in the correlation analysis presented in this paper, where the smallest effect was also found for the test that uses videos showing repairs to the hydraulic brake.

Further validity evidence can be drawn from a 2022 field study in which the "timing belt" vignette test was used in a real final exam ("Gesellenprüfung", $N = 156$ examinees). The video-based test was complemented by a hands-on exam station in which the apprentices had to perform maintenance work on an engine's timing chain. From a technical perspective, both systems are similar in their parts and function. However, the hands-on station used in our validation study—requiring working on the exact same engine type that is shown in the video vignettes—resembled the video-based test much more closely than the station used in the real exam. To our surprise, the correlation of our timing belt video test with the timing chain exam station was nominally higher ($\tau = 0.41$; $p_{\text{one-sided}} < 0.001$) than the correlation we found in our validation study ($\tau = 0.29$; $p_{\text{one-sided}} < 0.006$). This leads to the question of how motivated the participants of our validation study were. Most examinees *appeared* motivated; however, knowing that their performance was of no consequence to their school grades, and that no results would be made available to their teachers due to the anonymous character of the study, it is possible that they did not work as focused and as motivated as they would in their final exams.

Conclusions

It was the aim of this paper to evaluate and discuss validity evidence for a CBA that was designed as a complement of the practical part of the final exams for automotive technician apprentices in the German VET system. The focus of this study was on ecological validity, which can be subsumed under the concept of criterion-based validity.

Mixed conclusions can be drawn from the findings of this study. On one hand, our results prove that it is generally possible to design video-based items that predict the hands-on performance of technicians who carry out automotive maintenance and repair tasks. This finding is encouraging for the development of future CBAs that focus on hands-on skills. On the other hand, a good ecological validity was found only for a minority of the video-based items used in this study, and overall, the predictive potential of the three video-based instruments we investigated did not meet our expectations. Surprisingly, a better predictive potential was found when the same tool was put to practical use at a real journeyman's exam station.

From a test developer's perspective, the findings underline the importance of extensive validation. As Cook et al. (2013) and Gafni (2016) pointed out, many tests used in the VET sector lack sufficient validity evidence. Only if validity is systematically investigated by using evidence from various sources, can limitations in the appropriateness of test score interpretations such as those found in this study be uncovered and consequently eliminated. To achieve this, researchers must study the underlying causes for limited validity in future studies.

From a test user's perspective, this is especially important for exams, as the results have far-reaching consequences for the people tested.

A problem that emerged during the discussion of our findings is very limited research on the objectivity, reliability, and validity of the traditional assessment formats used in practical exams today is available. The almost complete lack of evidence about the psychometric properties of hands-on tests with human scoring inevitably leads to another question: How reliable and valid are computer-based assessments that *perfectly resemble* traditional exams if the measurement precision and appropriateness of those exams are unknown? At the worst, the use of tests with high ecological validity could result in one imprecise instrument replacing another. To improve job-related tests and exams, standards of educational testing must be applied to exams of any format, and they also must be communicated to and discussed with practitioners and decision takers throughout the vocational education, training, and assessment world.

Acknowledgements

The authors like to thank Stefan Müllerschön and his team at BiA Stuttgart (Bildungsakademie der Handwerkskammer Region Stuttgart), who made this study possible by providing equipment, workshop space, and professional experts who contributed as examiners.

Author contributions

All authors contributed substantially to this work. TG raised funding and coordinates the project. SH developed the theoretical framework in consultation with TG. Study design and data analysis for this paper were carried out by SH. EG supervised the study on location and contributed as an examiner. All authors discussed together the manuscript at all stages. All authors read and approved the final manuscript.

Funding

This study is part of project ***DigiDIn-Kfz*** of the research initiative ***ASCOT+***, funded by the German Federal Ministry of Education and Research, Grant no. 21AP004A. The funding body did in no way interfere with the design of the study, the collection, analysis, and interpretation of data, or the writing of the manuscript.

Availability of data and materials

The data and material of this study are not yet available to the public. In accordance with the regulations of the funding body, the data will be made available at the end of the project in 2023. The video-based tests used in this study are currently used in the official journeymen's exams at different locations in Germany. In order to prevent cheating or "teaching to the test", the test items cannot be made publicly available. However we included an example item in the article, hoping it will help readers to understand the general nature of the test.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 27 October 2022 Accepted: 7 November 2023
Published online: 06 December 2023

References

American Educational Research Association, & National Council on Measurement in Education [AERA, APA, & NCME] (2014) Standards for educational and psychological testing. American Educational Research Association, Washington, D.C.

Bejar II, Williamson DM, Mislevy RJ (2006) Human scoring. In: Williamson DM, Bejar II, Mislevy RJ (eds) Automated scoring of complex tasks in computer-based testing. Lawrence Erlbaum, Mahwah, pp 49–81

Bennett RE (2002) Inexorable and inevitable: the continuing story of technology and assessment. J Technol Learn Assess 1(1). http://www.jtla.org

Bennett RE, Braswell J, Oranje A, Sandene B, Kaplan B, Yan F (2008) Does it matter if i take my mathematics test on computer? A second empirical study of mode effects in NAEP. J Technol Learn Assess 6(9). http://www.jtla.org

Clariana R, Wallace P (2002) Paper-based versus computer-based assessment: key factors associated with the test mode effect. Br J Edu Technol 33(5):593–602. https://doi.org/10.1111/1467-8535.00294

Conole G, Warburton B (2005) A review of computer-assisted assessment. Res Learn Technol 13(1):17. https://doi.org/10.3402/rlt.v13i1.10970

Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R (2013) Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med 88(6):872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf

Deutscher Industrie- und Handelskammertag (2021). Wirtschaftspolitische Positionen der IHK-Organisation 2021 [Economic policy positions of the IHK organisation]. https://www.dihk.de/resource/blob/68502/e08ce6e2433b83ac5df677860c47b173/wirtschaftspolitische-positionen-2021-gesamtdokument-data.pdf. Accessed 24 Nov 2023

European Association for Research in Learning and Instruction [EARLI] (2023) EARLI 2023 book of abstracts. EARLI, Thessaloniki

Gafni N (2016) Comments on implementing validity theory. Assess Educ Princ Pol Pract 23(2):284–286. https://doi.org/10.1080/0969594X.2015.1111195

Gouvier W, Barker A, Musso M (2010) Ecological validity. In: Salkind NJ (ed) Encyclopedia of research design. SAGE, London, pp 399–404. https://doi.org/10.4135/9781412961288.n126

Gschwendtner T, Abele S, Nickolaus R (2009) Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern [Computer-simulated work samples: a validation study using the example of automotive technicians' fault diagnosis performance]. Zeitschrift für Berufs- und Wirtschaftspädagogik 105(4):557–578. https://doi.org/10.25162/zbw-2009-0038

Gschwendtner T, Abele S, Schmidt T, Nickolaus R (2017) Multidimensional competency assessments and structures in VET. In: Leutner D, Fleischer J, Grünkorn J, Klieme E (eds) Competence assessment in education. Research, models and instruments. Springer, Berlin, pp 183–202

Hartmann S, Gschwendtner T. (2021) Known-Groups-Validierung eines digitalen Prüfungsinstruments für Kfz-Mechatroniker*innen [Known-groups validation of a digital exam for light vehicle technicians]. Paper presented at the annual conference of the Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaften (DGfE), Bamberg

Hartmann S (2013) Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz [The role of reading comprehension and reading speed in text-based assessments of scientific inquiry skills] (Doctoral dissertation, University of Duisburg-Essen, Essen, Germany). https://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-33260/hartmann_diss.pdf. Accessed 24 Nov 2023

Hillen MA, van Vliet LM, de Haes HCJM, Smets EMA (2013) Developing and administering scripted video vignettes for experimental research of patient–provider communication. Patient Educ Couns 91(3):295–309. https://doi.org/10.1016/j.pec.2013.01.020

Holleman GA, Hooge ITC, Kemner C, Hessels RS (2020) The 'real-world approach' and its problems: a critique of the term ecological validity. Front Psychol 11:721. https://doi.org/10.3389/fpsyg.2020.00721

Kane MT (2016) Explicating validity. Assess Educ Prin Policy Pract 23(2):198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kingstone A, Smilek D, Eastwood JD (2008) Cognitive ethology: a new approach for studying human cognition. Br J Psychol 99:317–340. https://doi.org/10.1348/000712607X251243
Kirschner PA, Park B, Malone S, Jarodzka H (2017) Towards a cognitive theory of multimedia assessment (CTMMA). In: Spector JM, Lockee BB, Childress MD (eds) Learning, design, and technology: an international compendium of theory, research, practice, and policy. Springer, Cham, pp 1–23. https://doi.org/10.1007/978-3-319-17727-4_53-1
Knievel I, Lindmeier AM, Heinze A (2015) Beyond knowledge: measuring primary teachers' subject-specific competences in and for teaching mathematics with items based on video vignettes. Int J Sci Math Educ 13:1–21. https://doi.org/10.1007/s10763-014-9608-z
Malone S (2020) Technologiegestütztes assessment, online assessment [Technology-based assessment, online assessment]. In: Niegemann H, Weinberger A (eds) *Handbuch Bildungstechnologie* [Handbook of educational technology]. Springer, Berlin, pp 493–513
Messick S (1987) Validity (ETS Research Report No. RR-87–40). Educational Testing Service, Princeton
Moon JA, Keehner M, Katz IR (2019) Affordances of item formats and their effects on test-taker cognition under uncertainty. Educ Meas Issues Pract 38(1):54–62. https://doi.org/10.1111/emip.12229
Norwig K, Güzel E, Hartmann S, Gschwendtner T (2021) Tools to tap into the content of human minds": Think-Aloud-Interviews und Cognitive Labs als zentrale Bausteine zur Identifikation von Barrieren in Fehlerdiagnoseprozessen bei Auszubildenden des Kfz-Handwerks und zur Entwicklung adressatenspezifischer Lehr-/Lernarrangements [Think-aloud interviews and cognitive labs as central elements to identify cognitive barriers during diagnose processes in order to develop target-specific learning arrangements for light vehicle technicians]. Zeitschrift für Berufs- und Wirtschaftspädagogik 17(4):658–693. https://doi.org/10.25162/zbw-2021-0025
Palm T (2008) Performance assessment and authentic assessment: a conceptual analysis of the literature. Pract Assess Res Eval 13:4. https://doi.org/10.7275/0qpc-ws45
Parshall CG, Spray JA, Kalohn JC, Davey T (2002) Practical considerations in computer-based testing. Springer, New York. https://doi.org/10.1007/978-1-4613-0083-0
Parshall CG, Harmes JC, Davey T, Pashley PJ (2010) Innovative item types for computerized testing. In: van der Linden WJ, Glas CAW (eds) Elements of adaptive testing. Springer, New York, pp 215–230
R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Rausch A, Seifried J, Wuttke E, Kögler K, Brandt S (2016) Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. Empir Res Vocat Educ Train 8:9. https://doi.org/10.1186/s40461-016-0035-y
Revelle W (2020) psych: procedures for personality and psychological research. Northwestern University, Evanston
Rohr-Mentele S, Forster-Heinzer S (2021) Practical validation framework for competence measurement in VET: a validation study of an instrument for measuring basic commercial knowledge and skills in Switzerland. Empir Res Vocat Educ Train 13:18. https://doi.org/10.1186/s40461-021-00122-2
Rupp AA, Pant HA (2006) Validity theory. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. SAGE publications, Thousand Oaks, pp 1032–1035
Sadler PM (1998) Psychometric models of student conceptions in science: reconciling qualitative studies and distractor-driven assessment instruments. J Res Sci Teach 35:265–296
Spöttl G, Becker M, Musekamp F (2011) Anforderungen an Kfz-Mechatroniker und Implikationen für die Kompetenzerfassung [Requirements for automotive technicians, and implications on job assessment]. In: Nickolaus R, Pätzold G (eds) Lehr-Lernforschung in der gewerblich-technischen Berufsbildung, vol ZBW-Beiheft 25. Franz Steiner Verlag, Stuttgart, pp 37–53
Stecher BM, Klein SP (1997) The cost of science performance assessments in large-scale testing programs. Educ Eval Policy Anal 19:1–14
Stieler JF (2011) Validität summativer Prüfungen: Überlegungen zur Gestaltung von Klausuren [The validity of summative assessments: considerations on the design of exams]. Janus Presse, Bielefeld
The Commission on Technology and Adult Learning [CTAL] (2001) A vision of e-learning for America's workforce: report of the commission on technology and adult learning. https://web.archive.org/web/20030821165057if_/http://www.astd.org:80/virtual_community/public_policy/jh_ver.pdf. Accessed 24 Nov 2023
The International Test Commission [ITC] (2006) International guidelines on computer-based and internet-delivered testing. Int J Test 6(2):143–171. https://doi.org/10.1207/s15327574ijt0602_4
The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany [KMK] (2009) *Rahmenlehrplan für den Ausbildungsberuf Kraftfahrzeugmechatroniker und Kraftfahrzeugmechatronikerin* [Framework curriculum for the vocational training of automotive technicians]. KMK, Berlin
The transport training board (2023) Light vehicle technician apprenticeship. https://www.transporttraining.org/motor-industry/apprenticeship/light-vehicle-technician-apprenticeship-transport-training-services/
Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Venables B (2022) gplots: Various R programming tools for plotting data. https://CRAN.R-project.org/package=gplots. Accessed 24 Nov 2023
Weber W, Schmidt T, Abele S, Heilig S, Sarnitz A, Nickolaus R (2015) Kompetenzzuschreibungen von Ausbildern: Analyse zur Güte von Ausbilderurteilen [Competence attributions of trainers: Analysis of the quality of trainer judgements]. Zeitschrift für Berufs- und Wirtschaftspädagogik 111(1):125–136. https://doi.org/10.25162/zbw-2015-0007
Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York
Williamson DM, Bejar II, Mislevy RJ (2006) Automated scoring of complex tasks in computer-based testing: an introduction. In: Williamson DM, Bejar II, Mislevy RJ (eds) Automated scoring of complex tasks in computer-based testing. Lawrence Erlbaum, Mahwah, pp 1–13

Publisher's Note

Stefan Hartmann is an educational psychologist. His academic interests are in the field of research methodology and quantitative methods. A strong focus of his work is on quantitative assessments of skills and competences, and on the validity of the inferences drawn from the results of such assessments.

Emre Güzel is a technical instructor. His research interests are the development of tools to measure and promote vocational skills and competences. In his current project focuses on the development and evaluation of an interactive multimedia learning platform to promote troubleshooting skills of apprentice light vehicle technicians.

Tobias Gschwendtner is a Professor of technology and its didactics. His academic interests are in the field of digital diagnostics and intervention in automotive technology (VET) and technology education at secondary level 1.