Journal of Technical Education

Band 4, 2016, Heft 2

Stephan Abele (University of Stuttgart)

Can diagnostic problem-solving competences of car mechatronics be validly assessed using a paper-pencil test?

Journal of Technical Education (JOTED)

ISSN 2198-0306

Online unter: http://www.journal-of-technical-education.de

Stephan Abele (University of Stuttgart)

Can diagnostic problem-solving competences of car mechatronics be validly assessed using a paper-pencil test?

Abstract

In this study, an authentic paper-based key-feature test for electrotechnical diagnostic problem-solving competence was developed, administered to car mechatronic apprentices (N = 206) and validated using diagnostic problem-solving scores. It was hypothesized that the paper-based test scores represent the same construct as the problem-solving scores. The written test turned out to have a relatively high reliability (EAP/PV = .75). Additionally, it was found that the paper-based scores and problem-solving scores were highly correlated (r = .76) but represented empirically distinct dimensions. Presumably, the written test especially covered diagnostic knowledge and failed to cover other relevant subcomponents of diagnostic problem-solving competence. It is argued that this unexpected finding might be caused by construct underrepresentation and construct-irrelevant variance of the paper-based key-feature test.

Keywords: Professional competence, domain-specific problem-solving, key-feature items, construct representation, construct-irrelevant variance

Lassen sich diagnostische Problemlösekompetenzen von Kfz-Mechatronikern mit einem Papier-Bleistift-Test valide erfassen?

Zusammenfassung

In der Studie wurde ein schriftlicher Key-Feature-Test zur Erfassung diagnostischer Problemlösekompetenzen entwickelt, bei einer Stichprobe von Auszubildenden der Kfz-Mechatronik (N = 206) eingesetzt und anhand von Scores zum diagnostischen Problemlösen validiert. Es wurde unterstellt, dass die Papier-Bleistift-basierten Testscores dieselbe empirische Dimension abbilden wie die Problemlösescores. Der schriftliche Test erreichte eine relativ hohe Reliabilität (EAP/PV = .75). Gezeigt hat sich zudem, dass die Test- und Problemlösescores eng korrelierten (r = .76), letztlich aber empirisch unterscheidbare Dimensionen repräsentierten. Der schriftliche Test erfasste vermutlich v.a. diagnostisches Wissen und „vernachlässigte“ weitere relevante Subkomponenten der diagnostischen Problemlösekompetenz. Dieses unerwartete Ergebnis lässt sich wohl darauf zurückführen, dass der schriftliche Key-Feature-Test das Zielkonstrukt nicht vollständig repräsentierte und zudem Konstrukt-irrelevante Varianz erzeugte.

Schlüsselwörter: Berufsfachliche Kompetenz, domänenspezifisches Problemlösen, Key-Feature-Items, Konstruktrepräsentation, Konstrukt-irrelevante Varianz

1 Introduction

This paper is about the assessment of diagnostic problem-solving competences of car mechatronics[1]. Diagnostic problem-solving represents an important type of problem-solving (cf. Jonassen 2000) and is relevant in several professional contexts: Physicians have to identify causes of diseases, teachers have to figure out reasons for learning difficulties, technicians and engineers have to diagnose technical defects, etc. According to Shavelson (2010, p. 45), diagnostic problem-solving competence can be defined as a complex ability necessary to master challenging and somehow novel encounters occurring in the workplace. Diagnostic problem-solving competence is a key aspect of professional competence, a relevant subdimension of professional problem-solving competence and an important topic of Vocational Education and Training (e.g., Baethge & Arends 2009). The assessment of diagnostic problem-solving competence is a precondition of the evaluation of corresponding training programs and evidence-based feedback for students.

According to the current state of research, it seems advisable to assess diagnostic problem-solving competences using computer-simulations representing (parts of) the real work environment (e.g., Norcini & McKinley 2007): Such simulations allow for a high degree of validity, standardization and, in comparison to work-based assessments, are much more convenient. Although previous studies on professional problem-solving have documented the immense potential of computer-simulation-based testing (e.g., Rausch et al. 2016; Walker, Link & Nickolaus 2016; Gschwendtner, Abele & Nickolaus 2009), they have also made clear that such assessments imply long development times, high investment costs and also, due to time-consuming introduction into the computer-simulated work environments, high testing times. In contrast, paper-and-pencil-based tests are relatively economical and there are some findings suggesting that such tests might be appropriate to validly measure professional problem-solving competences (e.g., Link & Geißel, 2015). The present paper investigates whether diagnostic problem-solving competences of car mechatronics can be validly assessed using a paper-and-pencil-based (paper-based) test.

1.1 Diagnostic problem-solving competence of car mechatronics

1.1.1 Problem domain: Electrotechnical diagnostic problems

Diagnostic problems of car mechatronics refer to situations in which the cause(s) of an auto defect (e.g., lighting system defect) has/have to be found. Such situations have the critical attributes of a problem (cf. Jonassen 2000, p. 65): there is an unknown (e.g., cause of a lighting system defect) and it is worth finding this unknown (e.g., to satisfy a customer). This definition of a diagnostic problem is in line with Schaafstal, Schraagen & van Berlo (2000, p. 75) but differs from other studies (e.g., Kassirer, Wong & Kopelman 2010, p. 6) since it does *not* cover repair or maintenance.

[1] The field of car mechatronics covers, among other things, troubleshooting, repair and maintenance of cars (Baethge and Arends 2009).

Previous studies showed that professional problem-solving competence is domain-specific (e.g., van der Vleuten et al. 2010, p. 704). Without any further effort, high performance in a specific problem domain is not necessarily applicable to other problem domains (e.g., Schwartz & Elstein 2011, p. 225). There is evidence documenting the influence of complex (general) problem solving competence on professional competence being quite small (cf. Mainert et al. 2015), especially in comparison to the influence of domain-specific knowledge (cf. Abele et al. 2012). Thus, evidence suggests focusing on a specific problem domain when diagnostic problem-solving competences are measured. There is little consensus, however, on what exactly constitutes a problem domain and how to define it (e.g., Beck 2005, p. 551).

Electrotechnical diagnostic problems of car mechatronics are considered in this work. Diagnostic problems of car mechatronics can be subdivided into mechanical, electrotechnical and electro-mechanical problems. The domain of electrotechnical diagnostic problems consists of problems which require detecting electrotechnical causes of an auto defect. Such causes usually refer to broken electrotechnical components (actuators, sensors, control units, electric motors, lamps, fuses, etc.) and broken wires. Taking into account the international comparative analysis of Baethge and Arends (2009), such problems are (very) important transnational problems in the field of car mechatronics.

1.1.2 Electrotechnical diagnostic problem-solving competence

Electrotechnical diagnostic problem-solving competence is defined as the mental basis necessary to identify the electrotechnical cause(s) of an auto defect. It is assumed that this competence comprises several *subcomponents* and refers to successfully organizing a *process*.

Applying the model of scientific problem-solving of Klahr (2000), three sub-processes of the diagnostic problem-solving process can be distinguished: *hypothesis formulation*, *hypothesis testing* and *evidence evaluation*. This process structure has proved to be fruitful in many studies on diagnostic problem solving (e.g., Schaafstal, Schraagen & van Berlo, 2000, pp. 78–79). There are also findings suggesting that *information gathering* is another relevant sub-process (e.g., Roberts, While & Fitzpatrick, 1996).

Information gathering relates to collecting information about the symptoms of the car's defect(s), the technical particularities of the car (e.g., brand or vintage), relevant car systems (e.g., using technical information material), etc. In many cases, the diagnostic problem-solving process is supported by computer-based expert systems that provide, among other things, circuit diagrams and information on the location of car components. The results of the information gathering sub-process form the basis for formulating hypotheses on causes of the fault (*hypothesis formulation*). Sometimes the expert system can be used to read out the error-storage, which often indicates the "problem area", narrows down the space of possible causes and (sometimes) provides defect hypotheses. In the next step, the most probable hypothesis is tested (*hypothesis testing*), usually applying electronic test equipment (e.g., multimeters or oscilloscopes). Finally, the test result has to be evaluated (*evidence evaluation*). If the evaluation does not support the hypothesis, another hypothesis has to be tested, etc. Commonly, the problem-solving process includes several iterations and the order of the sub-processes deviates from the ideal process described here.

The outlined process is illustrated by an example. For illustration purposes, a relatively simple example is used. Facing a malfunctioning air conditioning unit, the technical particularities of the car (e.g.: What kind of air conditioning does the car have?) and the symptoms (e.g.: Is it possible to regulate the temperature?) of the defect have to be determined (*information gathering*). Afterwards, the computer-based expert system could be used to get details on the defect. These details could indicate that the malfunction is related to a specific sensor of the air conditioning system. It could be hypothesized that the sensor is defective, or that the wire connecting the relevant control unit and the air conditioning sensor is broken, etc. (*hypothesis formulation*). The hypothesis of a defective sensor might be tested by an electrotechnical measurement: the measurement of resistance (*hypothesis testing*). If the resistance of the sensor is infinite, the hypothesis is confirmed (*hypothesis evaluation*), i.e., a defective sensor is the cause of the malfunctioning air conditioning. The replacement of the air conditioning sensor is not part of the diagnostic problem-solving, but of the repair. If the resistance value lies in the target area, another hypothesis has to be generated, tested, and so on.

There have been ample studies showing that *diagnostic problem-solving knowledge* is a key subcomponent of diagnostic problem-solving competence (e.g., Walker, Link & Nickolaus 2016; Nickolaus et al. 2012). According to Jonassen and Hung (2006), diagnostic problem-solving knowledge is understood here as an integrative construct made up of several aspects: general domain knowledge (e.g., knowledge of Ohm's law), system knowledge (e.g., knowledge of the structure and function of electrotechnical car systems), device knowledge (e.g., knowledge of electronic test equipment) and experiential knowledge (e.g., historical knowledge of common causes of faults). Strategic knowledge, information gathering skills and measurement skills are other relevant aspects (e.g., Jonassen & Hung 2006). The *strategic knowledge* represents knowledge on how to coordinate and monitor the entire problem-solving process: it is, for example, the basis for decisions on which process step and sub-process should follow another or which hypothesis should be tested first. *Information gathering skills* are, among other things, required to use the computer-based expert system to read out the error-storage, to retrieve circuit diagrams and information on the location of car components. *Measurement skills* are necessary to apply electronic test equipment, i.e. to test error hypotheses. The development, mental transformation and application of this knowledge to specific diagnostic problems depend on *working memory capacity* as well as *causal and analytical reasoning*. It seems obvious that *motivational, volitional and emotional aspects* play a decisive role, too (cf. Sembill, Rausch & Kögler 2013).

2 Relation of diagnostic problem-solving competence and relevant paper-based tests

In the study of Walker, Link & Nickolaus (2016), 46 % of the variance of diagnostic problem-solving competence of electronics technicians was explained by a paper-based test score representing domain-specific knowledge. Abele (2014) found a correlation of r=.63 between a paper-based test for diagnostic knowledge and electrotechnical diagnostic problem-solving competence of car mechatronics. Nickolaus et al. (2012) presented an even higher correlation

of r = .80 between these measures. It is worth mentioning that these high correlations were achieved even though the paper-pencil tests were not explicitly developed to assess diagnostic problem-solving competence.

The review article of Swanson, Norcini and Grosso (1987) documents the correlation between paper-pencil tests and professional problem-solving competence being low to moderate, but the authors stress that these correlations might be heavily biased by the low reliabilities of the measures. When corrected for reliability, the correlations between paper-based scores and problem-solving competence might be very high (up to r = 1.0).

Link and Geißel (2015) showed that a specific professional problem-solving competence of electronics technicians[2] can be validly assessed using a paper-pencil test. In this study, the requirements of the paper-based test largely reflected the "reality", i.e., the authentic requirements occurring in the workplace. Authenticity seems to be one of the key aspects of the validity of an assessment in professional contexts (cf. Tigelaar & van der Vleuten 2014, p. 1251).

So we can see that there is some evidence that paper-based tests might validly represent diagnostic problem-solving competence in case the paper-based test is aligned to assess diagnostic problem-solving competence and has a high degree of authenticity and reliability. Even if the paper-based test does not completely cover the construct, the paper-based scores can provide a convenient basis to estimate the diagnostic problem-solving competence (i.e., the interindividual differences in competence) given the paper-based and the problem-solving scores are highly correlated (close to r = 1).

3 Paper-based key-feature tests: A promising method to assess electrotechnical diagnostic problem-solving competence

A very prominent paper-based method to assess diagnostic problem-solving competences is written simulations (cf. Norcini & McKinley 2007). In such assessments, the test takers are typically confronted with an authentic patient problem and they are asked to answer several questions referring to different aspects of one patient problem (cf. van der Vleuten 1996, p. 44). It is well known that the psychometric quality of the scores resulting from a one-problem-testing is commonly low (e.g., Greiff 2012, p. 72) – especially in terms of reliability. Additionally, such assessments require a comprehensive written introduction to, and explanation of, the patient's situation; so reading competence plays an important role and the assessment is probably biased by construct-irrelevant variance (cf. Kane 2013).

The paper-based *key-feature method* (cf. Hatala & Norman 2002) focuses on critical steps in solving diagnostic problems and provides the possibility, in comparison to the aforementioned method, to administer more and less text-laden items. Key-feature items consist of a short "stem followed by one or more questions" (cf. Fischer et al., 2005, p. 1) which allows several

[2] The study deals with constructive problem-solving competences, i.e., competences required to programming a programmable logic controller.

independent items to be administered. There is evidence that the key-feature approach can result in measures of good psychometric quality (cf. Hryncha, Takahash & Nayer 2014).

It has been well-documented (e.g., Klahr 2000) that successful diagnostic problem solvers are superior in formulating reasonable problem-specific hypotheses (*hypothesis formulation*) and in evaluating evidence resulting from hypothesis testing (*hypothesis evaluation*). Abele, Walker and Nickolaus (2014) showed that electrotechnical diagnostic problem-solving competence in car mechatronics can be validly measured with computer-based key-feature items that refer to using circuit and location diagrams (*information gathering*) and planning how to test diagnostic hypotheses (*hypothesis testing*). On the basis of a research review, Abele (2016) concluded that information gathering, hypothesis formulation, hypothesis testing and hypothesis evaluation are related to problem-solving activities that are critical to the diagnostic problem-solving success. Thus, it is advisable to develop key-feature items covering these critical sub-processes. It is sometimes argued that assessments concentrating on items covering specific aspects of a complex competence (i.e., key-feature items) run the serious risk of construct underrepresentation (e.g., Messick 1994).

In line with the MicroDyn approach (cf. Greiff 2012), it seems, however, possible to obtain an appropriate construct representation if there are sufficient key-feature items for each diagnostic problem-solving sub-process. MicroDyn was developed to assess complex problem-solving competence by using several items for each of the three theoretically distinguished complex problem-solving sub-processes. This approach turned out to be empirically fruitful (e.g., Greiff et al. 2013). Furthermore, there are findings proving that scores of relatively simple paper-based items and complex scenarios are closely correlated (r = .75-97: Swanson, Norcini and Grosso 1987, p. 236; Tigelaar & van der Vleuten 2014, p. 1244) and that complex problems scores "did not add much compared with relatively simple short…scenarios" (cf. van der Vleuten et al. 2010, p. 706).

In sum, it seems defendable to assess electrotechnical diagnostic problem-solving competence of car mechatronics using authentic key-feature items that draw on gathering information as well as formulating, testing and evaluating defect hypotheses. In order to achieve an acceptable construct representation, it is advisable to generate several key-feature items for each sub-process. Although the paper-based key-feature method is an integral part of medical assessment, this approach has not been used, to the best of my knowledge, in Vocational Education and Training to date.

4 Paper-based assessment of diagnostic problem-solving competence

The paper-based assessment was developed in line with the aforementioned argumentation. The assessment is comprised of four booklets: two information booklets containing circuit diagrams and diagrams for the location of car components, one booklet presenting the test stimuli and the questions (key-feature item booklet) and one booklet for the responses. The reason for developing four booklets was to come as close as possible to authentic diagnostic problem-solving. For example, having information booklets and an item booklet require

autonomous interactions to gather information depending on the actual status of the problem-solving process. Figure 1 gives screenshots of the information and item booklets.

Information booklet **Item stem (item booklet)**

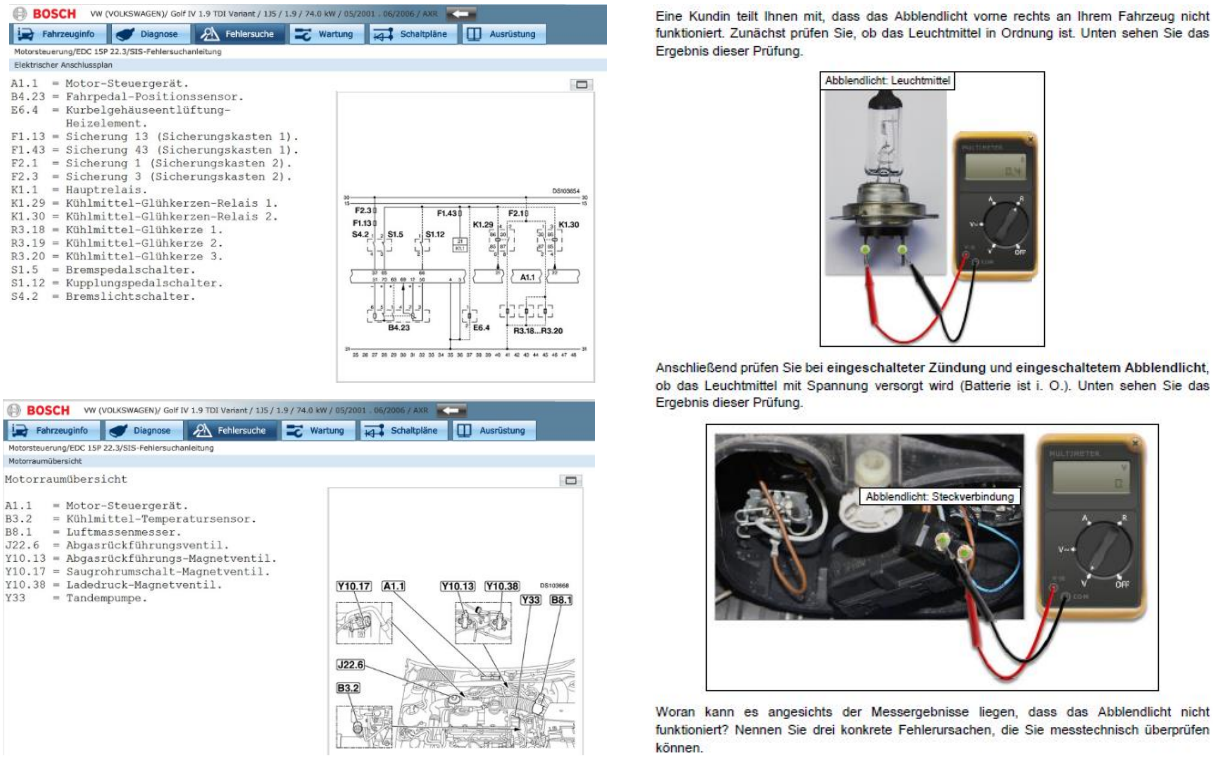

Figure 1: Examples of the information and item booklets (screenshots on the left: technical terms and abbreviated designation of car components, the corresponding circuit diagram and information on their location; screenshot on the right: the upper German text says that the customer complains about a faulty low beam headlamp. Furthermore, it says that the picture with the lamp and voltmeter shows the result of the first measurement. The text in the middle gives context information on the next measuring result, which is shown in the picture below. The text at the bottom includes the question: Considering the measuring results and the specific diagnostic situation, what are possible causes of the defective low beam headlamp? Please give three causes.)

The key-feature items were presented using authentic graphic material. Each item contained short textual descriptions, demanded interpretations of authentic visual stimulus material, referred to a specific problem-solving sub-process and encompassed several problem-solving steps. The amount of text, visual material and problem-solving steps varied between the items. The right part of Figure 2 gives an example of an “average” item. A lot of attention was paid to having only as much text as necessary. The descriptions are purposefully arranged to simulate the “real” sub-process as authentically as possible. Overall, two multiple-choice and 20 short-answer items were developed (Table 1).

ID	Aim and content	Sub-process	Car system
IG1	Localization of the speed sensor	Information gathering	Electronic engine management
IG2	Specifying measuring points at the intake-manifold change-over valve		
IG3	Localization of the exhaust-gas recirculation valve		
IG4	Specifying the abbreviated designation of the vehicle speed sensor		
IG5	Specifying the signal line of the camshaft sensor		
IG6	Specifying the purpose of a specific wire of the electric fuel pump		
IG7	Localization oft he pressure-charging valve		
IG8	Specifying the designation of the pressure-charging valve		
HF1	Formulation of hypotheses on causes of a defective fuel temperature sensor	Hypothesis formulation	Electronic engine management
HF2	Formulation of hypotheses on causes of a defective low beam headlamp		Lighting system
HF3	Formulation of hypotheses on causes of a defective spark plug relay		Electronic engine management
HF4	Formulation of hypotheses on causes of a defective engine control unit		Electronic engine management
HF5	Formulation of hypotheses on causes of a defective pressure-charging valve I		Electronic engine management
HF6	Formulation of hypotheses on causes of a defective pressure-charging valve II		Electronic engine management
HT1	Specifying a strategy to test a broken signal line of the camshaft sensor	Hypothesis testing	Electronic engine management
HT2	Selecting electronic test equipment to test the function of the electric starting system		Electronic engine management
HT3	Specifying a strategy to test a broken wire of the intake-manifold change-over valve		Electronic engine management
HT4	Specifying strategies to test hypotheses on a defective pressure-charging valve		Electronic engine management
EE1	Evaluation of measuring results and specifying the cause of a defective speed sensor	Evidence evaluation	Electronic engine management
EE2	Specifying the resistance value range of an intact fuse of the low beam headlamp		Lighting system
EE3	Specifying the pulse duty factor of the exhaust-gas recirculation valve using a signal measurement result of an oscilloscope		Electronic engine management
EE4	Evaluation of measuring results and specifying the cause of a defective diesel fuel injector unit		Electronic engine management

Table 1: Overview of the paper-based key-feature items

Several key-feature items (KFI) were developed for each diagnostic problem-solving sub-process. The *information gathering* KFI required specific information to be given by means of the information booklets (e.g., to find out the location of the speed and reference-mark sensor). The *hypothesis formulation* KFI demanded analysis of a diagnostic situation and formulation of error hypotheses (Figure 1). The *hypothesis testing* KFI were about strategies to test a specific hypothesis or select electronic test equipment in specific test situations. Finally, the *evidence evaluation* KFI demanded an interpretation of given measuring results. The 22 KFI cover the electronic engine management and lighting system.

Since we were aiming for high-quality items, we did not succeed in developing the same number of KFI for each sub-process: During the development, it turned out that it was easy to design authentic information gathering KFI and difficult to design authentic KFI on hypothesis testing and evidence evaluation, because some requirements were especially challenging to implement in a paper-based test. For example, hypothesis testing frequently is related to measurements, which cannot be considered in a paper-based test.

5 Research hypothesis

It is assumed that the paper-based test provides valid test score interpretations, meaning that the paper-based scores can be interpreted as indicators of diagnostic problem-solving competence. To test this hypothesis, the electrotechnical diagnostic problem-solving competence of car mechatronic apprentices was assessed and the outlined paper-based test was administered. In line with the hypothesis, evidence for convergent validity was expected, i.e., it was supposed that the paper-based scores and diagnostic problem-solving scores were unidimensional.

6 Method

6.1 Sample and design

In order to test the hypothesis, 206 car mechatronic apprentices nearing the end of the 3rd year of training were sampled from vocational schools of the federal state of Baden-Württemberg, Germany. The mean age of the apprentices was 21.0 years (SD = 3.1) and 6.2 % of the sample were females.

In the present study, a within-subjects-design was used. At the beginning, the apprentices were confronted with the paper-based test in the classroom. Approximately two weeks later, the electrotechnical diagnostic problem-solving competence was measured, administering 4 authentic diagnostic problems in a computer simulation, and 4 authentic diagnostic problems on a car in a garage. To control for position, i.e. exhaustion effects, the problems were administered in a *latin square design* (cf. Frey, Hartig & Rupp 2009, p. 45). For organizational reasons, it was not possible to control for order effects potentially caused by starting with the written test and ending with the authentic problems. The overall testing time

was 350 minutes (90 minutes for the paper-based test). Due to the long testing time, the written test was administered to a selection of apprentices (N = 121).

Prior to the written test, the sample got a standardized instruction of 10 minutes, including information on the testing time, the number of items, the test and response booklets as well as the information booklets. To prevent cheating, two tests were created, containing the same items in different orders.

The standardized instruction for the computer-based assessment took 30 minutes. The instructor demonstrated the handling of the computer simulation by means of a video projector. Afterwards, the apprentices worked individually on standardized tasks concerning dealing with the simulation. In very rare cases, apprentices could not complete a task. Then the instructor gave explanations in front of the class using the video projector. The introduction to the assessment in the garage took 15 minutes and explained the test setting. In both assessments, the apprentices were acquainted with the response booklet and got instructions on how to prepare the handwritten documentation. They were told to document each relevant problem-solving step (e.g., each electrotechnical measurement result) and to finish the documentation by giving a clear statement on the cause of the diagnostic problem.

6.2 Measures

6.2.2 Electrotechnical diagnostic problem-solving competence

The electrotechnical diagnostic problem-solving competence was measured using eight diagnostic problems (Table 2). The scores of the computer-based testing and of the testing on the car in the garage were highly correlated (r = .94, latent) providing strong evidence for convergent validity. This finding justified interpreting the scores as indicators of diagnostic problem-solving competence and integrating them into a single score (cf. Abele & Nickolaus 2016).

The scoring was conducted by two independent raters applying a coding manual to the handwritten documentation of the apprentices and resulted in both dichotomous and polytomous data. In very rare cases of diverging scores, content-oriented discussions produced consensual scoring. For example, a problem was considered solved if the correct cause of a problem (e.g., a broken central locking motor) had been identified, documented and proved by appropriate measurements. In case of correct solutions, the testee got the highest score. If relevant problem-solving steps were *not* documented, lower scores were assigned.

ID	Aim and content	Sub-process	Car system
P1	Diagnosing the cause of a defective window regulator lighting	Complete diagnostic problem-solving process	Comfort system: Door control
P2	Diagnosing the cause of a defective car radio		Comfort system: Car radio
P3	Diagnosing the cause of a defective air recirculation flap		Comfort system: Air conditioning
P4	Diagnosing the cause of a defective outflow-temperature sensor		Comfort system: Air conditioning
P5	Diagnosing the cause of a defective central locking motor		Comfort system: Door control
P6	Diagnosing the cause of a defective electric mirror adjustment		Comfort system: Door control
P7	Diagnosing the cause of a defective wheel speed sensor		Brake system
P8	Diagnosing the cause of a defective fresh air fan		Comfort system: Air conditioning

Table 2: Overview of the authentic diagnostic problems

6.2.2 Paper-and-pencil-based assessment

The paper-based assessment included 22 categorical items (Table 1). Compared to the aforementioned assessment, the key-feature method led to a greatly reduced testing time (from 260 to 90 minutes).

A manual including the correct answers was developed for the scoring. The KFI were scored based on a few keywords. Most of the KFI were dichotomous, although some were polytomous. Regarding the example in Figure 2, each correct cause gives one point, implying a maximum score of three.

6.3 Statistical procedures

6.3.1 Missing data

The paper-based items were administered to 121 of the 206 apprentices. Taking 121 apprentices as reference, 0 to 8.3 % of the item data was missing (item-level missingness), whereby most of the rates of missingness were 0 or very close to 0 (Table 4). Taking the complete sample of 206 apprentices as a reference, the rates of missingness of the problem-solving scores ranged from 11 to 14 % (Table 3).

Obviously, the missing data was mainly due to a somehow “planned” missing data design (person-level missingness). Thus, large parts of the missing data can be classified as missing completely at random (cf. Enders 2010, p. 22) meaning that the missing data should not bias the statistical parameter estimates (cf. Graham 2009, p. 553). Since the hypothesis was tested

using the polytomous Rasch model, the missing data did not lead to a listwise deletion of cases and all item responses could be included to test the hypothesis (cf. Boone, Staver & Yale 2014, p. 380) increasing the statistical power.

6.3.2 Analysis of KFI and diagnostic problems

To evaluate the psychometric properties of the KFI and the diagnostic problems, statistical procedures of the classical test theory (CTT) and item response theory (IRT) were carried out. In terms of CTT, the item discrimination, the relative frequency of correct responses (item difficulty) and Cohen's kappa (inter-rater reliability) were determined. For the estimation of the IRT parameters, the partial credit model of the polytomous Rasch family was applied (cf. Ostini, Finkelman & Nering 2013) generating item difficulties and corresponding estimation errors expressed in the logit metric. In order to evaluate whether the items fit the partial credit model, the item infit (weighted mean square statistic) and outfit (unweighted mean square statistic) were examined (cf. Bond & Fox 2007, p. 137). According to Wilson (2005), infit and outfit statistics ranging from 0.75 to 1.33 indicate an acceptable fit (p. 129). Bond and Fox (2007) point out that *t* statistics of the item infit and outfit that are higher than 2 and smaller than -2 indicate a significant misfit (p. 134). Since the *t* statistic, however, is fairly sample size sensitive, items were only dropped if the infit/outfit statistics and the related *t* statistics deviated from the above-mentioned target ranges (cf. Wilson 2005, p. 129). Finally, the EAP/PV reliability was estimated for both assessments, which can be interpreted similarly to Cronbach's alpha.

6.3.3 Testing and construct validity

The hypothesis of convergent validity was examined in two steps: In the first step, a one-dimensional partial credit model was estimated integrating the diagnostic problems and the paper-based KFI. In the second step, a two-dimensional model was estimated, where the problems and the KFI constitute separate dimensions. Afterwards, the correlation of the two dimensions was inspected and the following analyses were conducted to decide which model fit the data better: First, a chi-square difference test was performed (cf. Reckase 2009, p. 218) with a significance level of 1%. Second, the deviance, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) of the models were compared, where smaller values indicate a better model fit. AIC differences of 10 indicate serious model differences (cf. Burnham & Anderson 2004, p. 271). If the dividend of sample size and model parameter is smaller than 40, also using the CAIC is recommended (cf. Burnham & Anderson 2004, p. 270). The analyses were performed with the ConQuest software (cf. Wu et al. 2007).

7 Results

7.1 Problem and KFI statistics

Table 3 gives the CTT statistics of the problems and the corresponding parameters of the partial credit model and documents the fact that the psychometric quality of the problems was good. Only the inter-rater reliability of problem P8 was remarkably low. Since the other

coefficients of P8 were acceptable, this problem was not dropped. The EAP/PV reliability of the measurement is .65, suggesting that the number of problems should be increased in future studies (e.g., Norcini & McKinley 2007, p. 243). It should be noted that, for parameterization reasons, the difficulties of the highest problem scores were constrained, so there were no standard errors available.

ID	κ	*f*	r	Difficulty	Error	Outfit	Outfit t	Infit	Infit t	N
P1.1	.91	.22	.61	.50	.18	.98	-.2	.99	-.1	184
P1.2		.35		-.25	constrained	.92	-.7	.92	-1.1	
P2.1	.93	.20	.55	.28	.19	.96	-.3	.98	-.1	183
P2.2		.47		-.70	constrained	.96	-.4	.98	-.3	
P3.1	.96	.35	.62	-.66	.16	.96	-.4	.97	-.5	182
P3.2		.31		.19	.17	.98	-.2	.99	-.1	
P3.3		.12		1.40	constrained	.88	-1.1	.95	-.3	
P4.1	.83	.16	.56	-.10	.20	.92	-.7	.98	-.1	183
P4.2		.63		-1.32	constrained	.88	-1.2	.92	-1.2	
P5.1	.97	.25	.46	.98	.18	.97	-.2	1.00	.0	180
P5.2		.09		1.43	constrained	.90	-1.0	1.04	.3	
P6.1	.94	.20	.49	.54	.19	1.04	.4	1.02	.2	177
P6.2		.40		-.54	constrained	1.13	1.2	1.11	1.6	
P7.1	.92	.08	.60	1.67	.27	1.17	1.6	1.01	.1	185
P7.2		.42		-1.46	constrained	.99	-.1	.99	-.1	
P8	.69	.87	.38	-2.02	.23	.87	-1.3	.95	-.3	179

Notes. κ = inter-rater reliability; *f* = relative frequency of correct responses per category, the frequency of incorrect responses results from 1 minus the sum of the relative frequencies; *r* = discrimination: Pearson correlation between the item score and the total raw score; Difficulty = IRT difficulty (logit); error = standard error related to the difficulty estimation.

Table 3: CTT statistics and parameters of the partial credit model of the diagnostic problems

Table 4 gives the statistics of the paper-based KFI. Here, the discrimination of IG7 and IG8 as well as the Kappa of IG3 are considerably low. But again, these items were not excluded from further analyses due to the good quality of their other statistics. The EAP/PV reliability of the paper-based scores was .75.

ID	κ	*f*	r	Difficulty	Error	Outfit	Outfit t	Infit	Infit t	N
IG1	.97	.42	.27	.39	.20	1.09	.7	1.10	1.6	120
IG2	.85	.95	.35	-3.19	.43	1.09	.8	.93	-.1	120
IG3	.49	.42	.23	.37	.20	1.13	1.0	1.12	1.9	119
IG4	1.0	.88	.25	-2.24	.29	.94	-.4	1.00	.0	121
IG5	.92	.71	.41	-1.00	.21	1.04	.4	.98	-.2	121
IG6	.94	.80	.47	-1.55	.24	.82	-1.5	.89	-.8	121
IG7	.93	.78	.15	-1.39	.23	1.06	.5	1.05	.4	121
IG8	.97	.95	.17	-3.21	.43	.83	-1.3	1.01	.2	121
HF1.1	.89	.31	.71	-.33	.20	.95	-.3	.96	-.5	111
HF1.2		.27		.25	.22	.93	-.5	1.03	-.3	
HF1.3		.13		1.25	constrained	.58	-3.7	.91	-.4	
HF2.1	.89	.20	.61	-1.41	.21	1.01	.1	1.01	.1	111
HF2.2		.49		-1.05	.19	1.00	.0	1.00	.1	
HF2.3		.23		1.05	constrained	.83	-1.3	.97	-.2	
HF3.1	.92	.40	.60	-.61	.20	.96	-.2	.97	-.7	111
HF3.2		.32		.40	constrained	.84	-1.2	.93	-.8	
HF4	.81	.78	.50	-1.45	.24	.78	-1.7	.89	-.9	111
HF5.1	.86	.28	.53	-.74	.21	.97	-.2	.99	-.1	111
HF5.2		.52		-.55	constrained	.87	-.9	.91	-1.1	
HF6.1	.84	.37	.59	-1.35	.19	.94	-.4	.96	-.7	120
HF6.2		.49		-.22	constrained	.82	-1.4	.87	-2.0	
HT1	.97	.73	.39	-1.09	.22	.92	-.6	.95	-.5	121
HT2	.97	.55	.16	-.24	.19	1.14	1.1	1.14	2.2	121
HT3	.94	.20	.20	1.55	.25	1.11	.9	1.11	.8	111
HT4	.92	.38	.43	.54	.20	.92	-.6	.96	-.5	120
EE1	.52	.66	.33	-.74	.20	.94	-.4	.97	-.4	121
EE2	.96	.73	.35	-1.09	.22	1.06	.5	.98	-.1	121
EE3	.97	.63	.20	-.58	.20	1.10	.8	1.07	1.0	121
EE4	.61	.51	.54	-.05	.19	.87	-1.0	.89	-2.0	121

Notes. κ = inter-rater reliability; *f* = relative frequency of correct responses per category, the frequency of incorrect responses results from 1 minus the sum of the relative frequencies; *r* = discrimination: Pearson correlation between the item score and the total raw score; Difficulty = IRT difficulty (logit); error = standard error related to the difficulty estimation.

Table 4: CTT statistics and parameters of the partial credit model of the KFI

7.2 Evidence for the convergent validity

It was hypothesized that the paper-based scores and the problem-solving scores represent the same construct, so a one-dimensional model should fit the data at least as well as a two-dimensional model. In contrast to this hypothesis, the IRT analyses revealed that the two-dimensional partial credit model outperformed the unidimensional one (Table 5).

Model	N	*df*	Deviance	*BIC*	*AIC*	*CAIC*
One-dimensional	206	46	5,949.1	6,194.2	6,041.1	6,240.2
Two-dimensional	206	46	5,935.7	6,191.4	6,031.7	6,239.4

Table 5: Statistics of the one and two-dimensional partial credit model

The AIC documented the better fit regarding the information criteria of the two-dimensional compared to the one-dimensional model (Δ9.4). Moreover, the chi-square difference test was statistically significant (Δ*df* = 2; Δdeviance = 13.4; *p* = .001) indicating the superiority of the two-dimensional model. The correlation between the computer-based and paper-based scores was high (*r* = .76, latent) but substantially deviated from a perfect correlation.

8 Discussion

8.1 Summary

The aim of the present study was to investigate whether diagnostic problem-solving competences of car mechatronics can be validly measured using a paper-based test. For that purpose, a paper-based key-feature test for electrotechnical diagnostic problem-solving competence was developed, administered to a sample of car mechatronic apprentices and validated using diagnostic problem-solving scores. It was hypothesized that the paper-based scores would represent the same empirical dimension as the problem-solving scores, i.e., provide valid test score interpretations. Compared to previous studies (e.g., Hryncha, Takahash, & Nayer 2014), the key-feature test resulted in a relatively high reliability. Although the paper-based and problem-solving scores were highly correlated, there was, however, reasonable evidence for their discriminant validity: Model comparisons documented a superior fit of the model in which the paper-based scores and problem-solving scores represented distinct dimensions. Presumably, the written test especially tapped electrotechnical diagnostic knowledge which has been frequently proven to be distinguishable from, but closely correlated with, diagnostic problem-solving competence (e.g., Abele 2014, pp. 57–58; Nickolaus et al. 2012). So the paper-based scores reflected a key subcomponent of diagnostic problem-solving competence but failed to cover some other relevant subcomponents. In contrast to Link and Geißel (2015), the findings of this study suggest that paper-based tests are not completely suitable to measure professional problem-solving competence. What is the reason for these contradictory findings?

Content validation indicates that the paper-based key-feature test likely suffered *construct underrepresentation*: The written test did not tap an important subcomponent of electrotechnical diagnostic problem-solving competence, namely: measurement skills (i.e., conducting electrotechnical measurements). In a paper-based test, it seems (almost) impossible to authentically consider electrotechnical measurement skills of car mechatronics. Moreover, the key-feature items only covered diagnostic problem-solving sub-processes, i.e., strategic knowledge needed to coordinate the sub-processes and the entire diagnostic problem-solving process was not taken into account. Link and Geißel (2015) did not use key-

feature items to assess professional problem-solving competence but rather "holistic" items that covered the relevant subcomponents of their focal construct. Therefore, they probably succeeded in having an appropriate construct representation.

Abele, Walker and Nickolaus (2014) concluded that electrotechnical diagnostic problem-solving competence of car mechatronics can be validly measured using *computer-based key-feature* items. Here, it was concluded that the key-feature items are not completely suitable to measure professional problem-solving competence. There are at least three reasons for these somehow contradictory conclusions: (1) *statistical uncertainty*, (2) *construct under-representation* and (3) *construct-irrelevant variance*. (1) The results of Abele, Walker and Nickolaus (2014) were ambiguous. The statistical comparison of a two-dimensional model (dimension 1: computer-based key-feature scores; dimension 2: diagnostic problem-solving scores) to a unidimensional model did not bring clear results: The *p*-value of the model comparison was p=.04 indicating that the decision whether the two-dimensional or the unidimensional model fits the data better depends on which significance level is chosen (α=.01 or α=.05). Pragmatically, the authors argued that the latent correlation of r=.89 between the two dimensions is high enough to defend the use of computer-based key-feature items to measure diagnostic problem-solving competence. (2) Besides this statistical uncertainty, the contradictory findings might be caused by construct underrepresentation: In the study of Abele, Walker and Nickolaus (2014), the *computer-based* key-feature items had to be mastered by interacting with a computer-simulation reflecting parts of the real work environment of car mechatronics. In the *written* test, the "natural" work environment of car mechatronics could be only simulated rudimentarily. For example, information gathering using information booklets substantially differs from information gathering using the computer-based expert system: The paper-based key-feature items did not (authentically) tap information gathering skills. As mentioned before, the paper-based test did also not allow for including measurement skills, whereas the computer-based key-feature items did. (3) The paper-based test was probably also biased by somehow "artificial" interactions, i.e. construct-irrelevant variance: Although the development of information booklets aimed to come close to authentic information gathering interactions, there seem to be remarkable differences between the interactions connected to the paper-based test and the "reality".

Since the findings of this paper are bound to the domain of "electrotechnical diagnostic problems" of car mechatronics, the following question arises: What can we learn from this study about whether paper-based (key-feature) tests are appropriate to measure professional problem-solving competences?

8.2 Theoretical and practical significance

In view of the study of Link and Geißel (2015), it seems scientifically unjustifiable to principally consider *paper-based* tests as inappropriate to measure professional problem-solving competence, as might be concluded on the basis of other publications (e.g., van der Vleuten et al. 2010, p. 706). The decisive question is whether a test allows an adequate representation of the focal construct and causes construct-irrelevant variance (e.g., Messick 1994). For example, it could be appropriate to measure specific professional problem-solving

competences (e.g., designing electric circuits) using paper-based items. There should be, however, many more professional situations in which paper-based tests fall short of completely covering the professional problem-solving competence.

What about *key-feature* items? Generally speaking, authentic "holistic" items should imply a better construct representation and less construct-irrelevant variance than paper-based key-feature items. It should be pointed out, however, that "holistic" items can cause a severe amount of construct-irrelevant variance too – for example, when they require long textual introductions that are not associated with the focal construct. In the end, the question whether, and in which situations and domains, *key-feature* items are appropriate to measure professional problem-solving competences seems to still be open. Undoubtedly, the answer to this question relies on whether rather pragmatic arguments are accepted: In some contexts, latent correlations of $r \approx .80$ might be sufficient to take key-feature scores as acceptable proxies, i.e. valid indicators of professional problem-solving competence.

The arguments presented indicate that construct representation and construct-irrelevant variance are crucial to validation studies. This spotlights the role of the definition of professional problem-solving competence, which is the basis for judging whether a test is "contaminated" by construct underrepresentation and construct-irrelevant variance. This point is touched on in the *Standards for Educational and Psychological Testing* (cf. AERA, APA & NCME 2014), which highlights the fact that an important concern of validity is the use and interpretation of test scores (cf. AERA, APA & NCME 2014, p. 11). I think it is worth putting more emphasis on the construct definition. The definition of the problem-solving competence (the problem domain, subcomponents of problem-solving competence and problem-solving process) is the decisive reference point of the test score interpretation. It is the theoretical anchor of the validation process. Focusing more on the definition of professional problem-solving competence brings major advantages: It forces scientists to carefully describe and examine the professional problem-solving competence under investigation, increases our knowledge of professional problem-solving competence (e.g., its subcomponents and processes) as well as the precision, practical relevance and testability of corresponding theories.

Practically speaking, the domain-specificity of professional problem-solving competence is (very) challenging: Assuming that there are numerous professional problem domains, numerous different tests of professional problem-solving competence must be developed. There are findings showing that, within one occupational profile, the domain-specific problem-solving competences can be strongly correlated (e.g., $r = .77$; Walker, Link & Nickolaus 2016). These results could be used to legitimatize the generalization of domain-specific test results to other problem domains within one occupational profile. This argument might enormously reduce the number of assessments which are necessary to measure professional problem-solving competences. Its quality, however, strongly depends on the learning opportunities of the testees: If they did not have the opportunity to acquire the diverse domain-specific professional problem-solving competences related to an occupational profile, the generalization fails. Thus, the previous learning experiences of the testees and the organization of their vocational education should be accounted for when domain-specific

problem-solving scores are generalized. It might be fruitful to scrutinize the generalizability of domain-specific professional problem-solving scores (cf. Kane 2013) and to find out which measurement error is to be expected when professional problems of different problem domains are sampled to assess the problem-solving competence within one occupational profile. If these studies reveal that, under certain conditions, it is defensible to infer the overall problem-solving competence within one occupational profile from the scores of one or a selection of problem domains, enormous progress would be achieved.

8.3 Limitations and outlook

The empirical results of this study rest upon relatively small samples (N = 119-185). Penfield (2013) argues that "a sample size of 100 can be viewed as a lower threshold" (p. 132) for classical test theory analyses. Following Boone, Staver and Yale (2014), a sample size of 100 might be also sufficient to apply the Partial Credit Model (p. 364). Against this background, it can be assumed that the sample size might not decisively bias the insights presented here.

Another limitation is that the unidimensionality of the paper-based key-feature items (KFI) was not examined, even though the KFI only covered sub-processes. A supplementary confirmatory factor analysis revealed that the item scores referring to different sub-processes can be treated as unidimensional (N = 121, χ^2 = 247, df = 260.5, χ^2/df = 1.05, CFI = .95, $NNFI$ = .94, $WRMR$ = .95). It should also be noted that the sample size of this confirmatory analysis is relatively small.

As mentioned before, this study is limited to a specific problem domain. Further studies should examine whether the argumentation of this study can be confirmed in other problem domains. Furthermore, studies examining the professional problem-solving process and its mental subcomponents can bring substantial progress and enhance the understanding of professional problem-solving competence. In order to study the problem-solving process, computer-generated log-file data offer interesting options (cf. Greiff et al. 2013).

9 References

Abele, S. (2014). Modellierung und Entwicklung berufsfachlicher Kompetenz in der gewerblich-technischen Ausbildung. [Modeling and development of professional competence in Vocational Technical Education]. Stuttgart: Franz Steiner.

Abele, S. (2016). Theory of the Diagnostic Problem-solving Process in Professional Contexts and its Evaluation Using Computer-generated Log-file Data. Manuscript submitted for publication.

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R. & Funke, J. (2012). Dynamische Problemlösekompetenz – ein bedeutsamer Prädiktor von Problemlöseleistungen in technischen Anforderungskontexten? [Dynamic problem solving – an important predictor of problem-solving performance in technical domains?]. Zeitschrift für Erziehungswissenschaft, 15, 363–391.

Abele, S. & Nickolaus, R. (2016). Validität einer Diagnostik berufsfachlicher Problemlösekompetenzen bei Kfz-Mechatronikern mit einer Computersimulation [Validity of a computer-simulation-based assessment of professional problem-solving competence in the domain of car mechatronics]. Manuscript in preparation.

Abele, S., Walker, F. & Nickolaus, R. (2014). Zeitökonomische und reliable Diagnostik beruflicher Problemlösekompetenzen bei Auszubildenden zum Kfz-Mechatroniker [Time-saving and reliable diagnostics in measuring professional problem-solving competence in the domain of car mechatronics]. Zeitschrift für Pädagogische Psychologie, 28(4), 167–179.

AERA (American Education Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education) (2014). Standards for Educational and Psychological Testing. Washington D.C.: American Educational Research Association.

Baethge, M. & Arends, L. (2009). Feasibility study VET-LSA: A comparative analysis of occupational profiles and VET programmes in 8 European countries - International report. Vocational Training Research: Vol. 8. Bonn: Federal Ministry of Education and Research 22 (BMBF).

Beck, K. (2005). Ergebnisse und Desiderate zur Lehr-Lern-Forschung in der kaufmännischen Berufsausbildung [Results and desiderata of research on teaching-learning-processes in the field of commercial vocational education and training]. Zeitschrift für Berufs- und Wirtschaftspädagogik, 101(4), 533–556.

Bond, T. G. & Fox, C. M. (2007). Applying the Rasch Model: Fundamental measurement in the Human Sciences. Mahwah, NJ: Erlbaum.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). Rasch analysis in the human sciences. Dordrecht: Springer.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research, 33(2).

Enders, C. K. (2010). Applied missing data analysis. New York, NY: Guilford Press.

Fischer, M. R., Kopp, V., Holzer, M., Ruderich, F. & Jünger, J. (2005). A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Medical Teacher, 27, 1–6.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. Educational Measurement: Issues and Practice, 28(3), 39–53.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. Annual Review of Psychology, 60, 549–576.

Greiff, S. (2012). Individualdiagnostik komplexer Problemlösefähigkeit. Münster: Waxmann.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F. & Funke, J. (2013). Computer-based assessment of Complex Problem Solving: concept, implementation, and application. Educational Technology Research and Development, 61, 407–421.

Gschwendtner, T., Abele, S. & Nickolaus, R. (2009). Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern [Computer-simulation-based work samples: A validation study taking troubleshooting performance of car mechatronics as an example]. Zeitschrift für Berufs- und Wirtschaftspädagogik, 105, 557–578.

Hatala, R. & Norman, G. R. (2002). Adapting the key features examination for a clinical clerkship. Medical Education, 36, 160–165.

Hryncha, P., Takahash, S. G. & Nayer, M. (2014). Key-feature questions for assessment of clinical reasoning: a literature review. Medical Education, 48, 870–883.

Jonassen, D. H. (2000). Toward a Design Theory of Problem Solving. Educational Technology Research and Development, 48(4), 63–85.

Jonassen, D. H. & Hung, W. (2006). Learning to troubleshoot: A new theory-based design architecture. Educational Psychology Review, 18(1), 77–114.

Kane, M. J. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Kassirer, J., Wong, J. & Kopelman, R. (2010). Learning clinical reasoning (3rd ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Klahr, D. (2000). Scientific discovery as problem solving. In D. Klahr (Ed.), Exploring science. The cognition and development of discovery processes (pp. 21–39). Cambridge, MA: MIT Press.

Link, N. & Geißel, B. (2015) Konstruktvalidität konstruktiver Problemlösefähigkeit bei Elektronikern für Automatisierungstechnik. [Construct validity of the constructive problem-solving competence of electronics technicians for automation technology]. Zeitschrift für Berufs- und Wirtschaftspädagogik, 111, 208–221.

Mainert, J., Kretzschmar, A., Neubert, J. C. & Greiff, S. (2015). Linking complex problem solving and general mental ability to career advancement: Does a transversal skill reveal incremental predictive validity? International Journal of Lifelong Education, 34(4).

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13–23.

Nickolaus, R., Abele, S., Gschwendtner, T., Nitzschke, A. & Greiff, S. (2012). Fachspezifische Problemlösefähigkeit in gewerblich technischen Ausbildungsberufen – Modellierung, erreichte Niveaus und relevante Einflussfaktoren. [Domain-specific problem-solving competence in industrial-technical professions – Modelling, achieved levels and relevant predictors]. Zeitschrift für Berufs- und Wirtschaftspädagogik, 108, 243–272.

Norcini, J. J. & McKinley, D. W. (2007). Assessment methods in medical education. Teaching and Teacher Education, 23, 239–250.

Ostini, R., Finkelman, M. & Nering, M. (2013). Selecting among polytomous IRT models. In S. P. Reise & D. A. Revicki (Eds.), Handbook of item response theory modeling: Applications to typical performance assessment (pp. 285–304). New York, NY: Routledge.

Penfield, R. D. (2013). Item analysis. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise & M. C. Rodriguez (Eds.), Test theory and testing and assessment in industrial and organizational psychology: Vol. 1. APA handbook of testing and assessment in psychology (pp. 121–138). Washington, DC: American Psychological Association.

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. Empirical Research in Vocational Education and Training, 8(1), 1–23. doi:10.1186/s40461-016-0035-y

Reckase, M. D. (2009). Multidimensional item response theory. Dordrecht: Springer.

Roberts, J., While, A. E. & Fitzpatrick, J. (1996). Exploring the process of data acquisition: Methodological challenges encountered and strategies employed. Journal of Advanced Nursing, 23, 366–372.

Schaafstal, A., Schraagen, J. M. & van Berlo, M. (2000). Cognitive task analysis and innovation of training: the case of structured troubleshooting. Human Factors, 42.

Schwartz, A. & Elstein, A. S. (2011). Clinical reasoning in medicine. In J. Higgs, M. A. Jones, S. Loftus & N. Christensen (Eds.), Clinical Reasoning in the Health Professions (3rd ed., pp. 223–234). Oxford: Elsevier Ltd.

Sembill, D., Rausch, A. & Kögler, K. (2013). Non-cognitive facets of competence: Theoretical foundations and implications for measurement. In O. Zlatkin-Troitschanskaia & K. Beck (Eds.), From Diagnostics to learning success - Proceedings in vocational education and training. Rotterdam: Sense.

Shavelson, R. J. (2010). On the measurement of competency. Empirical Research in Vocational Education and Training, 2(1), 41–63.

Swanson, D. B., Norcini, J. J. & Grosso, L. J. (1987). Assessment of clinical competence: Written and computer-based simulations. Assessment and Evaluation in Higher Education, 12(3).

Tigelaar, D. E. & van der Vleuten, C. P. (2014). Assessment of professional competence. In S. Billet, C. Harteis & H. Gruber (Eds.), International Handbook on Research in Professional and Practice-based Learning (pp. 1237–1270). Dordrecht: Springer.

van der Vleuten, C. (1996). The assessment of professional competence: Developments research and practical implications. Advances in Health Science Education, 1, 41–67.

van der Vleuten, C. P., Schuwirth, L. W., Scheele, F., Driessen, E. W. & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. Best Practice & Research Clinical Obstetrics and Gynaecology, 24, 703–719.

Walker, F., Link, N. & Nickolaus, R. (2016). A multidimensional structure of problem-solving competencies of electronics technicians for automation technology. Empirical Research in Vocational Education and Training, 8, 1-16. doi:10.1186/s40461-016-0034-z

Wilson, M. (2005). Constructing Measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER ConQuest Version 2.0: Generalized item response modelling software. Camberwell, VIC: ACER Press.

Author

Dr. Stephan Abele

Department of Vocational Education, Institute for Educational Science, University of Stuttgart

Geschwister-Scholl-Straße 24D, 70174 Stuttgart, Germany

abele@bwt.uni-stuttgart.de

Zitieren dieses Beitrages:

Abele, S. (2016). Can diagnostic problem-solving competencies of car mechatronics be validly assessed using a paper-pencil test? Journal of Technical Education (JOTED), Jg. 4 (Heft 2), S. 190-211.