GUIDELINES FOR COMPETENCE ASSESSMENT IN ENGINEERING EDUCATION AN IMPLEMENTATION IN PROJECT NUSAL

Albers, Albert; Butenko, Viktoriia; Breitschuh, Jan; Walter, Benjamin; Drechsler, Sandra; Burkardt, Norbert
Karlsruhe Institute of Technology (KIT), Germany

Abstract
Methods of competence modelling and assessment used in psychological research hold high potential for gaining insight into requirements to engineers in product design. The contribution describes methods for competence modelling, task and test development and test analysis and optimization.
The methods describes are exemplarily implemented in a federal funded project about Standardization in Higher Education in Germany and the imtermediate results in this ongoing work are discussed.

Keywords: Design Education, Human behaviour in design, Research methodologies and methods, Evaluation

Contact:
Prof. Dr.-Ing. Albert Albers
Karlsruhe Institute of Technology (KIT)
IPEK Institute of Product Engineering
Germany
albert.albers@kit.edu

Please cite this paper as:
Surnames, Initials: *Title of paper*. In: Proceedings of the 20th International Conference on Engineering Design (ICED15), Vol. nn: Title of Volume, Milan, Italy, 27.-30.07.2015

1 INTRODUCTION

Mechanical Engineering is a highly requested course in German tertiary education resulting in a high number of enlisted students. Unfortunately, universities face an increasing shortage of available resources. In order to ensure high quality of teaching it is thus necessary to spend as much resources on the actual learning process of the students as possible. Consequently, the effort invested in the accompanying processes of teaching such as assessments of learning success should be minimized (Rossiter, 2013). On the other hand, high quality assessments provide deep insight into the students' actual competence profiles (Biggs and Tang, 2007). This contradiction makes the use of efficient methods for competence assessment a vital part of modern higher education.

The German Federal Ministry of Education and Research (BMBF) has made several attempts over the recent years to strengthen professional higher education by funding of multiple research projects (Koeppen et. al., 2008, Weishaupt et. al., 2012). One of the largest projects is "Competence Modelling and Competence Assessment in Higher Education KoKoHs" (Zlatkin-Troitschanskaia, 2013). Within this project a test instrument for the domain of Engineering Mechanics has been developed, implemented and optimized (Musekamp et. al., 2013). Due to this experience the BMBF and the German Institute for Standardization (DIN) entrusted the IPEK – Institute of Product Engineering with the development of a competence model for the whole domain of standardization across major disciplines such as engineering, healthcare, financial services or church organisations within the project "NuSaL – Standardization in Higher Education".

In this work general requirements for the development of competence models and corresponding tests will be outlined from an engineering point of view, to give a link between the psychological aspects of testing and the current practices in engineering tests (e.g. exams). It is to be noted, that "test" may be interpreted as either written or oral assessment; the quality criteria remain the same. What changes with the type of assessment, is the step in the assessment process, in which the criteria manifest most. While in written assessment task and test design are crucial, in oral assessments aspects such as matching between different examiners and consistency of inter-personal relationships gain relevance (Joughin, 1998). The examples given in this work will focus aspects of written assessments.

Section 2 explains the steps of competence modelling, test concept development, task construction, pilot phase and task selection and optimization by giving requirements for the respective steps and outlining some of the most important associated tools and methods. Section 3 shows how these principals were implemented in "NuSaL" and Section 4 gives an intermediate conclusion and outlook.

2 GENERAL STEPS AND REQUIREMENTS OF TEST DEVELOPMENT

2.1 Modelling Competences

Competences in the context of this work are regarded as learnable, context specific performance dispositions according to Weinert (2002, p. 17-31). Competence models thus describe the elementary abilities and their connections in a hierarchical way by means of competence dimensions. The quantity of dimensions is individual and different for each special field. A dimension can be divided into sub-dimensions, the quantity of which may also be different. Each sub-dimension corresponds to a predefined ability according to the particular higher dimension. The competence dimensions can be isolated from each other, as well as hierarchical related to each other. This heavily depends on the field.

Generic competence models such as e.g. Bloom's Taxonomy (Krathwohl, 2002) can be seen as frameworks for domain-specific competence models; each vocational domain (e.g. mechanical engineering, medical sciences or financial services) has got specific demands for abilities, a graduate should possess. It is thus purposeful to consolidate the detailed cognitive processes from generic models into relevant competence dimensions for a specific field. This may lead to stronger focus on some cognitive processes (e.g. factual or procedural knowledge) in the domain-specific competence models.

Nevertheless, the competence dimensions should be defined in a way that they represent the theoretically and empirically verifiable basis for the development of test tasks (so called items). This means that the verification of the competence model is carried out by designing and implementing tests that explicitly match the model's assumptions and restrictions (see also section 2.2).

The competence dimensions should be defined unambiguous which can require the consolidation of multiple dimensions into one. On the other hand, a dimension should be divided into other sub-dimensions, if major parts of the associated competences are not logically or thematically connected to each other. All sub-dimensions should have an immediate relation to the particular higher dimension.
For the development of a competence model, it is first necessary to identify the core abilities and demands in the selected competence areas. This survey can happen through several means, including interviews, observations, questionnaire surveys, and research investigations. For qualitative empirical study with a great significance of practical experience, interviews play a key role as a good survey tool (Bühner, 2004, p. 101ff). If a systematic registration of the competence dimensions is done based on learning books, curricula, literature, special focus must be laid on covering a broad range of relevant input (Bühner, 2004, p. 100f). Usually, a mix of both approaches holds good results.

2.2 Development of the Test Concept

A good test construct should fulfil the following main quality criteria: objectivity, reliability and validity.
Objectivity is understood as the independence of test results from the type of execution or the person executing. A distinction is made between three kinds of objectivity: objectivity of execution, evaluation and interpretation (Lienert and Raatz, 1998, p. 7f). The objectivity of execution is reached when both the behaviour of the test supervisor and the conditions of the execution have the least possible influence on test results. This can be achieved by high standardization of the test situation and a complete restriction of individual leeway during execution by using identical test material and standardized instructions. Another important quality criterion is the objectivity of the evaluation by using the same criteria for evaluating measurements that cannot be influenced by a person's subjective abilities and experience. Especially for manual evaluation, the objectivity of evaluation should be sophistically reasoned and the level of freedom regarding the evaluation of test results should be kept at a minimum. The final important criterion of objectivity is objectivity of interpretation. It embodies clearly defined rules that allow an identical interpretation of test values and the results based on the test person's individual position on respective competence dimensions. This results from the exact definition of abilities that are required for a certain level of competence. For a statistical approach every satisfied criterion is parameterized with a certain score (Lienert and Raatz, p. 147ff).
The second aforementioned quality criterion of a test construct is *reliability* (Lienert and Raatz, p. 173ff). It depicts the accuracy by which a test measures a certain competence dimension. A test is reliable if the localization of the subject on the respective competence dimensions occurs absolutely free of errors. This can be realized e.g. by a test repetition with an equal sample of subjects. Good reliability of the test construct can be assumed if there are no significant differences between the results of both tests.
The last criterion for test quality is *validity* (Lienert and Raatz, p. 220f). It indicates the level of accuracy by which the test measures the abilities being studied. Validity can be measured in a number of ways e.g. via subject proficient item (i.e. task) analysis or statistical sample validity measures.
The test construct can be implemented as "pen and paper" or "computer assisted" test. Computer assisted tests are more sustainable due to paperless testing materials and save time in evaluating test results. Additionally computer assisted tests hold easy possibility for good objectivity of rating and fairness in assessing each test subject as independence from the test supervisor is warranted (Bull and Danson, 2004). On the other hand, the development of a pen-and-paper test takes less time and is less complex compared to the computer-assisted solution. The same goes for the evaluation of open questions. Hence it should be thought about the items format and the pros and cons of each tool in advance of choosing the survey instrument. (Bühner, 2004 p. 108ff).

2.3 Development of Tasks

Items are a means of measurement for the abilities of the subjects in selected fields. If the items are not developed correctly and do not measure what they are meant to, the test construct will show a low validity. It is important to note, that items should be carefully designed to require one specific competence facet according to the desired model (compare chapter 2.1). Therefore the following rules should be considered in the development of the items.
The development of items is not only challenging and requires creativity as well as a strict compliance with many rules e.g. objectivity, validity of content, one-dimensionality, independence and economy

of evaluation (Lienert and Raatz, 1998 p. 29ff). Below, the individual rules are briefly outlined on a few examples.
Objectivity describes the comprehensibility of elaborated items. This means that the task definition shall be perfectly understood by all students equally fast (independent of the particular subject). The use of a simple language, several review loops and wording improvements help in achieving this goal.
In general, there are two major types of questions: The types of questions querying an open answer require autonomous knowledge reproduction and entail a significantly higher evaluation effort compared to questions with a restricted answer format (Bühner, 2004, S. 125ff). These can be designed as allocation or selection tasks. These kinds of items are mostly used for knowledge- and proficiency examinations as well as the correct recognition of cause- and effect relationships. They are characterized by clear, economical and objective evaluation but only register a recognition effort. A high guessing probability resulting in a distortion of the results might be the case. The chance probability can be significantly reduced via a higher number of distractors. It is very important to attend to the formulation of distractors carefully. It should be avoided to use different wording to express the same statement for two distractors.
As one of the important criteria of a testing tool, the tasks should show content validity, which means that the test tasks should measure exactly what is intended for. If for example the proficiency in the area 'History of Standardization' shall be measured, items should singularly focus on this subject matter and no other. Another aspect of quality is the unidimensionality of the items in a test, which means that an item is assigned to exactly one dimension of the competence model by means of contents. This can be accomplished if the answers of the tasks address exactly one ability. The development of such test constructs is not trivial, but only this way it is later on possible, to statistically verify the degree of fit of the assumed competence model to empirically gathered data. Furthermore it is important to pay attention to the independency of the items, because the solving probability of an item shall only depend on the item difficulty and not be based on the result of another item.

2.4 Pilot Study

The pilot study serves as an empirical revision and validation of the design of the competence model and the items. The important result of this study is a statement whether the competence model can empirically prove itself and competencies of subjects can be differentially described with the developed test instrument based on the postulated competence dimensions.
For the pilot study a manageable sample size that is not too large and not too small is recommended. Before the survey it is important to establish a sufficient test subject motivation as the emotional and motivational processes have a high impact on the performance during information processing. The background of the study should be outlined shortly and exact information concerning testing range, working time and mode of task completion should be given. Maintaining the same procedure for all surveys is important in order to guarantee a high objectivity of execution. As the execution of the test is restricted in time it might happen that the tasks are not entirely completed leading to the exclusion of tasks at the end of the testing sheets. In order to avoid omitting the same tasks over and over, different testing sheets should be developed and deployed (Hussy and Jain, 2002, p. 67ff). Every test sheet should have different variants in the sequence of items. A variation of the items can be achieved as follows: beginning with easy items increase the motivation of the subjects and complex items increase the concentration. At the end of a survey it is important to gather qualitative feedback from the subjects regarding the tasks to set impulses for a subsequent optimization and revision of items.

2.5 Task Selection and Optimization

Item evaluation and optimization are the final crucial steps before the main survey. The validity of the results in the subsequent main evaluation phase strongly depends on the empiric verification of the test draft version. This is why enough time should be planned for these activities. The main criteria to take into account during evaluation are outlined below.
Within the scope of item analysis the following criteria e.g. analysis of complexity and reliability are identified. The complexity analysis of items gives an overview on how difficult the item solving is for the test subjects. For the correct identification of the complexity index, omitted items are excluded in the evaluation. Otherwise they will lead to an over- or underestimation of the complexity that will

result in affecting both - reliability and validity of the test construct. The reliability analysis assesses the quality of the items and the entire scale.
The mode of calculation for the difficulty index strongly depends on the task format and the relating risk of guessing probability. For items with sufficient distractors, guessing plays a subordinate role and can therefore be ignored during calculation. Calculations are as follows:

$$P = \frac{Nr}{N} \times 100 \tag{1}$$

P = Difficulty index
Nr = Number of test subjects who solved problems correctly
N = Number of test subjects

The difficulty should be well distributed to match the estimated and observed distribution of abilities in the sample (Lienert and Raatz, 1998, p. 72ff).
Selectivity or discrimination of a task is the task's ability to differentiate between different levels of a persons abilities and is defined as the point-biserial correlation of an item with the scale without this item:

$$r_{i(x-i)} = \frac{\bar{X}_R - \bar{X}}{s_X} \sqrt{\frac{N_R}{N - N_R}} \tag{2}$$

$\bar{X}$ = Means of all raw scores
$\bar{X}_R$ = Means of raw score for subjects with correct answer
s_X = Standard deviation of raw scores
N = Number of test subjects
N_R = Number of test subjects with correct answers

There can be a total distinguish between three possible values, as shown in Figure 1.

Selectivity close 1.0	Selectivity close 0	Selectivity close -1.0
This item is just solved by test persons with high performance and is not solved by people with low performance	The item has no connection with differentiation by the total test and is not suitable to differentiate between subjects with high and low performance	The item is dissolved by test persons with low performance and not dissolved by the persons with better performance.

Figure 1. Possible extreme values of selectivity

Another important measure for task suitability is the internal consistency of scales. Usually, Cronbach's alpha is calculated for this purpose (Bühner, 2004, p. 166f). Cronbach's alpha is calculated as follows:

$$\alpha = \frac{k}{k-1} \cdot \left(1 - \frac{\sum_{i=1}^{k} {s_i}^2}{{s_x}^2}\right) \tag{3}$$

k = Number of items
s_i^2 = Variance of the item
s_x^2 = Variance of the scale

Cronbach's alpha allows for analysis whether or not the items of a scale purposefully measure the same construct. Nevertheless, its usefulness is the subject of on-going debates because of its easy misinterpretation as a measure for item homogeneity (Bühner, 2004, p. 168).

3 IMPLEMENTATION IN PROJECT NUSAL

3.1 Background

The project "Standardization in Higher Education (NuSaL)" is funded by the European Commission as well as the Federal Ministry for Education and Research (BMBF).
The goal is to refer to the relevance of standardization for innovation assurance, to strengthen the competitiveness of Germany and to adapt the academic education in accordance to future economic needs.

3.2 Competence Model

Basis for the development of the competence model in the area of standardization were requirements towards young professionals expected by the industry. A total of N=28 companies from different economic sectors were surveyed e.g. mechanical engineering, electrical engineering, IT, healthcare, financial services (banking and insurance) as well as church organisations. The selection of the interview partners was made with the aim to gather data from a large bandwidth of different industrial sectors in order to enable a statement concerning the relevance of the topic for different areas. 7 main dimensions could be derived from the identified requirements. These were subdivided into a total of 15 sub-dimensions, see Figure 2.

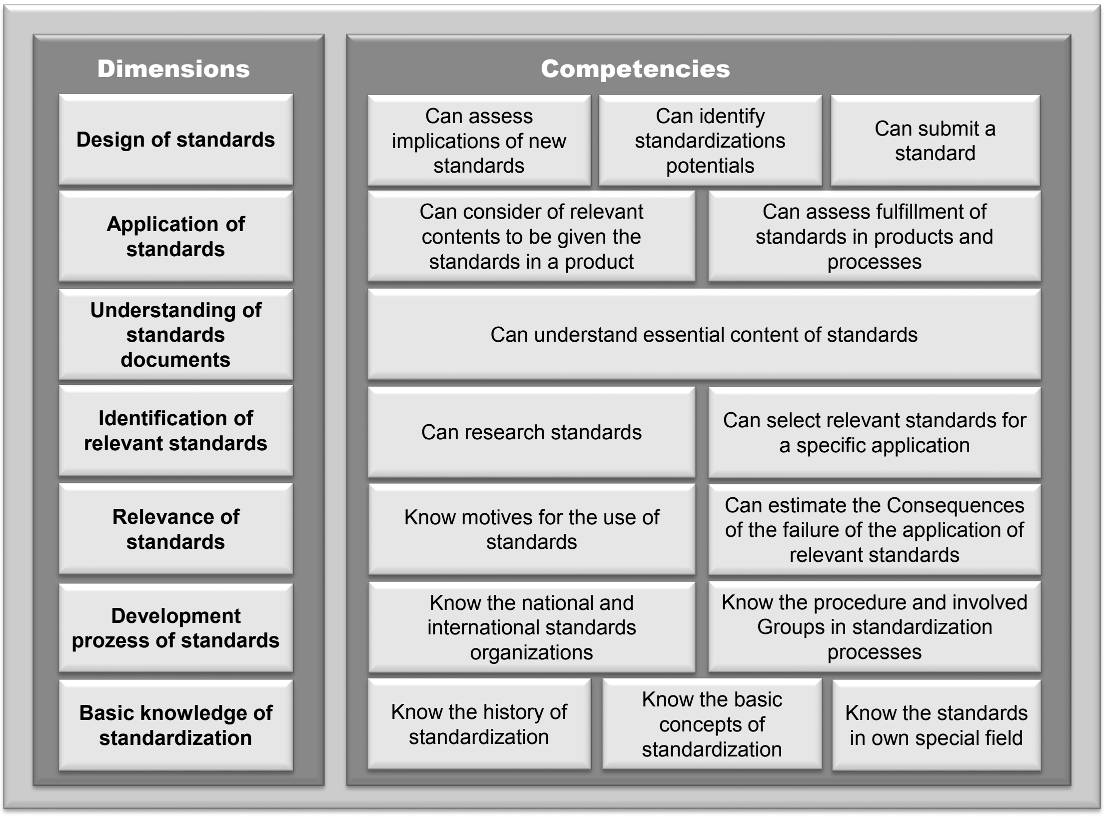

Figure 2. Competence model in project NuSaL

In order to be able to distinguish between the levels of knowledge of test subjects, every competence is differentiated into 4 competence characteristics: novice, beginner, intermediate and expert.
Every competence level describes exactly one skill that should be present. In Figure 3 all characteristics depending on the level of competence are indicated exemplarily by means of the sub-competence "knowledge of national and international normalization organizations".

Figure 3. Competence characteristics

3.3 Test Concept

The defined competence levels deliver a central aid for the item development. For the measurement of competences in scope of the NuSaL project, 38 items were compiled. Measurement of skills such as knowledge, comprehension and abilities has shown to be unproblematic and can be tested well via multiple choice and open questions. The last competence dimension "Design of Standards" is problematic. As it relates to working experience that can only be attained in industry (these abilities cannot be gained in theory) and are therefore not measurable in the scope of the project (which focusses on tertiary education). The number of items per dimension varies between 3 and 15 items and strongly depends on the number of sub-dimensions.

Within the project different courses such as mathematics, science, engineering, business and / economics were tested resulting in additional requirements concerning the test construct as every faculty has different study contents. Therefore the test instrument shall be developed universally in order to be applicable towards all areas of study. This was accomplished by designing closed questions for all essential subjects that are identical throughout the faculties. For example knowledge in the subject of identifying "National and International Normalization Organizations" is easily retrieved interdisciplinary via the use of multiple-choice questions. For subjects requiring a differentiated knowledge background, open questions were used.

In order to test the skill "Comprehension of Normalization- and Standardization Documents" in the faculties “mechanical engineering” and “biology”, different problem cases had to be developed since Standards have profoundly different vocabulary and therefore it is impossible to test all faculties with a given single scenario constructed problem.

3.4 Tasks

The developed test construct consists of a mix between multiple choice and open questions. For the development of multiple choice questions a significant emphasis was put on the formulation and optimum number of distractors (i.e. wrong answers) in order to minimize guessing probability. An exemplary multiple choice question (in this case a multiple-right-wrong selection) is shown in Figure 4. Scoring was done per row of the matrix, which allows analysing the item difficulty per sub-question.

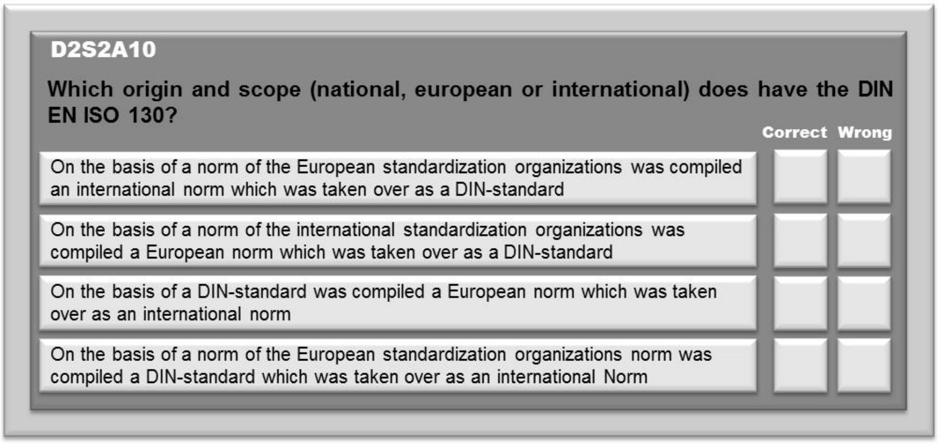

Figure 4. Multiple choice question

The development of open questions is focused on a simple and comprehensible formulation. An example for an open item is shown in Figure 5.

Figure 5. Open Question

3.5 Pilot Study

For the pilot study N=60 students participated from graduate studies "Mechanical Development" (N=23) and "Integrated Product Development" (N=37). The duration of the pilot study tests was 90 minutes. Both surveys had a standardized procedure. Initially the students were briefed on the project goal and motivated for participation accordingly. Thereinafter they were instructed in the problem format and the available time. In order to avoid systematic omission of items at the end of the survey, seven different item booklets were developed in which the tasks were listed in different order. A pen and paper approach was used as survey instrument in the pilot phase.

3.6 Task Selection and Optimization

Following the survey a correction of completed items was conducted, which resulted in 37 completed items. All evaluation criteria were defined beforehand for each item. In the course of the correction additionally required criteria were identified and existing criteria were adapted. Omitted items were not rated as zero, but marked so that they could be considered in later evaluation. The findings on which items were omitted mostly delivers very good indicators for the following revision of the problems and the targeted improvement of the test construct's quality criteria.

After the results of the correction phase were available they were analysed with the help of the Statistical Package for the Social Sciences Software (SPSS). Focus was put on item difficulties, item-scale-correlations as well as item separation. In Figure 6 the reliability statistics for 9 items from one scale is shown. The Cronbach’s alpha value reached 0.340. Values of 0.8 and above are considered acceptable. However, in practice usually much lower coefficient still are accepted.

Cronbach's Alpha	Cronbach's Alpha for standardized items	Quantity of items
,340	,341	9

Figure 6. Reliability Statistics

In figure 7 the item scale statistics that represent which items should be removed from the scale is shown, so that the homogeneity of the scale can be improved. The corrected item scale correlation is also known as selectivity of the items and indicates the correlation between an item and the scale without this item. The Cronbach's alpha column indicates whether the internal consistency of the scale can be improved if an item of little selectivity is removed from the scale.

	Average scale, if item deleted	Variance scale, if item deleted	Corrected item-scale correlation	Cronbach's Alpha if Item deleted
D2S2A1	,563	,087	,307	,216
D2S2A2	,538	,143	-,486	,575
D2S2A3	,575	,074	,562	,057
D2S2A4	,600	,091	,277	,239
D2S2A5	,600	,112	-,083	,412
D2S2A6	,612	,099	,166	,298
D2S2A7	,638	,103	,198	,294
D2S2A8	,612	,104	-069	,341
D2S2A9	,563	,081	,408	,155

Figure 7. Item Scale Statistics

The items D2S2A2, D2S2A5 and D2S2A8 have a negative selectivity, which indicates that people with high abilities score worse on these items than people with lower ability values which is a logical contradiction and is non-compliant with the competence model. This may be caused by faulty formulation of the items or misinterpretation by the test takers. However, if these three items are removed from the scale, scale quality can thus be improved noticeably (up to $\alpha = .575$).

4 INTERMEDIATE CONCLUSION AND OUTLOOK

The item selection resulted in an under-representation of certain scales. These scales (and the corresponding competence facets) must thus be collapsed in a way that complies with the competence model. This results in a reduced size of the competence model. Another way to fill the gap in data is to make new surveys with a larger sample size.

Nevertheless the competence model’s structure must show it holds true in the empirical data. For this purpose a factor analysis will be executed. Preliminary Rasch-analysis (Masters, 1982) showed partial compliance with the polytomous Rasch-model for some items. Further investigation in applying these models, which would allow for more sophisticated investigation of measurement quality, will be carried out.

Once test development is finished, investigation of the distribution of competences regarding standardization in German higher education will be done.

REFERENCES

Biggs, J. ., & Tang, C. (2007). Teaching for Quality Learning at University: what the student does. Maidenhead, U.K: Society for Research into Higher Education & Open University Press.

Bühner, M. (2004). Einführung in die Test- und Fragebogenkonstruktion. München [u.a.]: Pearson Studium.

Bull, J., & Danson, M. (2004). Computer-assisted assessment (CAA). Assessment Series, (14).

Hussy, W., & Jain, A. (2002). Experimentelle Hypothesenprüfung in der Psychologie. Göttingen: Hogrefe.

Joughin, G. (1998). Dimensions of Oral Assessment. Assessment & Evaluation in Higher Education 23(4): 367–378.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. Zeitschrift Für Psychologie / Journal of Psychology, 216(2), 61–73. doi:10.1027/0044-3409.216.2.61

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

Lienert, G. A., & Raatz, U. (1998). Testaufbau und Testanalyse. Weinheim: Beltz, Psychologie Verl.-Union.

Masters, G. N. (1982). A rasch model for partial credit scoring. Psychometrika, 47(2), 149–174. doi:10.1007/BF02296272

Moosbrugger, H., & Kelava, A. (2012). Testtheorie und Fragebogenkonstruktion. Berlin: Springer.

Musekamp, F., Mehrafza, M., Heine, J.-H., Schreiber, B., Saniter, A., Spöttl, G., Breitschuh, J., Heene, M., Butenko, V., Turki, T., Burkardt, N., Albers, A. (2013). Formatives Assessment fachlicher Kompetenzen von angehenden Ingenieuren – Validierung eines Kompetenzmodells für die Technische Mechanik im Inhaltsbereich Statik. In Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften (Vol. Sonderheft). Landau in der Pfalz: Verlag Empirische Pädagogik.

Rossiter, J. A. (2013). Case studies in making assessment efficient while developing student professionalism and managing transition. European Journal of Engineering Education, 38(6), 582–594. doi:10.1080/03043797.2013.824409

Weinert, F. E. (2002). Leistungsmessungen in Schulen. Weinheim [u.a.]: Beltz-Verl.

Weishaupt, H., Rittberger, M., Botte, A., Dees, W., Huth, R., Martini, R., & Muders, S. (2012). Bildungsforschung in Deutschland - eine Situationsanalyse. (H. Weishaupt & M. Rittberger, Eds.) (Vol. 37).

Zlatkin-Troitschanskaia, O., Reinhold, N., & Beck, K. (2013). Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften (Vol. Sonderheft). Landau in der Pfalz: Verlag Empirische Pädagogik.