INVESTIGATING LEARNING AND IMPROVING TEACHING IN ENGINEERING THERMODYNAMICS GUIDED BY CONSTRUCTIVE ALIGNMENT AND COMPETENCY MODELING:

PART II. ASSESSMENT & EXAM DESIGN

Thorsten Braun, Rolf Stierle, Matthias Fischer, and Joachim Gross
University of Stuttgart • Keplerstraße 7, 70174 Stuttgart, Germany

INTRODUCTION

In Part I of our series,[1] we proposed a competency model for our course on engineering thermodynamics, where we define competency as a holistic trait of the students who are able to successfully complete our one-year course. Furthermore, we show how we support student learning, especially through *cognitive apprenticeship*, transparent grading, and good solution manuals for the problem sets used in plenary exercises/guided tutorials/and especially in the final exams, etc. In addition to *generic math competency*, *subject specific problem solving*, and *subject knowledge*, we identified *exam proficiency* with respect to the final exam in our course as a professional competency considered important by the students. Therefore, we mean to answer the questions: *Do we test the competencies we want to convey to our students?* and *what are the actual, empirical competencies that students acquire throughout the one-year course on engineering thermodynamics?* We accomplish this by applying models of probabilistic item response theory to six of our exams between winter semester 2016/2017 and summer semester 2019 (for raw exam data and more detailed description of variables, consult Braun[2]). This analysis is driven by the desire to improve exam quality, align the exam with the learning objectives and learning activities as advocated by *constructive alignment* and contribute to a competency model for engineering thermodynamics (proposed in Part I of our series [1]). The following analysis contributes to such a model through a thorough investigation into the actual empirical dimensions of competency as measured by our exams. We will show that different thermodynamic subjects are less important for exam performance than we expected and that different dimensions of students' competency are observed *within* single exam tasks rather than *between* them. The results are related to recent theoretical and empirical discussions in engineering education research.

This work is structured as follows. We first present selected results from the probabilistic analysis of past exams to reveal the competency-structure of our final exams.

Thorsten Braun *works as a higher education developer and educational researcher at the Center for Higher Education and Lifelong Learning at the University of Stuttgart. He studied sociology, political science, and law at the Technical University of Darmstadt and the Philipps University of Marburg. In 2020, he completed his PhD thesis on influencing factors on study success, based on a mixed methods case study of a course in engineering thermodynamics. ORCID 0000-0002-8418-2520.*

Rolf Stierle *is a research associate in the group of Joachim Gross at the University of Stuttgart, where he also teaches a course on non-equilibrium thermodynamics. He studied chemical engineering at the University of Stuttgart and received his PhD for his work on non-equilibrium processes at vapor-liquid interfaces. His research interests include the incorporation of molecular detail into non-equilibrium models of interfacial processes based on classical density functional theory. ORCID: 0000-0001-7475-7207*

Matthias Fischer *is a former research assistant in the group of Joachim Gross at the University of Stuttgart. There he assisted in teaching and organizing numerous courses, including Engineering Thermodynamics 1/2. He studied chemical engineering at the University of Stuttgart and received his PhD in the area of molecular simulations for dynamic properties of fluids. He now works in industry on research and development of drying systems for agricultural products. ORCID: 0000-0002-9832-629X*

Joachim Gross *is Professor of Thermodynamics and Thermal Process Engineering and Dean of Chemical Engineering Studies at the University of Stuttgart. He received his chemical engineering education at the Technical University of Berlin and, after 4 years in industry, was appointed Professor of Separation Technology and later Engineering Thermodynamics at Delft University of Technology. His research focuses on prediction of physical properties based on molecular methods and non-equilibrium thermodynamics to improve the efficiency of thermal separation processes. ORCID: 0000-0001-8632-357X*

Next, we show how the continuing analysis led to quality guidelines for exam design that improved the overall test quality. By further combining the results from probabilistic modeling and additional qualitative survey data, we answer the question of what actually makes exam tasks in thermodynamics easy or difficult for students. These insights are of paramount importance for exam design, but also for the students themselves, who benefit from a generalized and empirical view on the exam as a challenging situation. The results help students to assess their own exam experiences and performance in comparison to their peers, which in turn supports individual reflection on learning strategies and learning progress. Finally, we conclude with an outlook on future measures related to the final exam to further improve our course on engineering thermodynamics.

EXAM ANALYSIS AND THE EMPIRICAL DIMENSIONS OF COMPETENCY

In this section we present the results of the probabilistic modeling analysis of our final exams. The purpose of the different modeling approaches is to contrast our expected dimensions of competency, as presented in Part I of our series,[1] with the empirical, actual dimensions as represented by students' exam responses. We define competency as a holistic trait of students who are able to successfully complete our one-year course on engineering thermodynamics. Dealing with this situation does not only include the exam-performance; the necessary competency also extends to the learning environment, personal and social conditions, and so on. [1]

Modeling Approach

A suitable way to learn about students' competency is to measure their performance. This assumes that a psychometric test or an exam represents a latent trait or a person's ability to perform a specific competency. In our case we thoroughly analyzed students' responses to our exams from 2015 to 2019 in order to discover the actual competency-structure hidden within the exam results. The number of students in every exam analyzed is given in Table 1.

A common approach to evaluating exam results is a summation (cumulative grading), and perhaps some form of basic distribution, of students' achieved points compared to their expected levels of performance. This kind of simple analysis has some shortcomings in that it does not represent the empirical difficulty of exam tasks in relation to the students' performance, which basically means high measurement error and very rudimentary information on how suitable the exam tasks are for measuring the students' level of proficiency. We try to overcome this problem by applying psychometric approaches of exam modeling. While these are more complex, they provide significantly enhanced insight into the underlying structure of student responses.

TABLE 1
Number of students who participated in the exam for *Engineering Thermodynamics 1/2* in recent years. This equals the case numbers *N* for the different item response theory models.

Semester	Number of Students
Summer 2015	322
Winter 2016/17	203
Summer 2017	360
Winter 2017/18	192
Summer 2018	333
Winter 2018/19	206
Summer 2019	262

The exam analysis in this work is conducted using probabilistic test theory; in particular we use *item response theory* (IRT) based on the Rasch model and familiar logistic variants.[3] Because of the way the exam was designed, we apply three different major models to the data (the students' responses to the exam). First, we apply a uni-dimensional Rasch model (one-parameter logistic model, 1PL),[4,5] which assumes that our exam represents a holistic thermodynamic competency without significant sub-dimensions; second, a Rasch testlet bifactor model (BF) [6,7] that expects students to have one general and one or more subject-specific competencies; and third, a partial credit model (PCM) [8-10] that shifts the attention to distinct sub-competencies *within* individual exam tasks.

All calculations are performed with the *test analysis module* (TAM) package for the statistical software *R*.[11] Some statistics for model fit comparison and model quality have been summarized in Table 2. Essentially, the statistical models are an assessment of the difficulty of an exam task relative to all other tasks in the exam by calculating the most likely distribution of exam results that satisfy the assumptions of the model. This means creating a data matrix of the exam results with each exam task and every student's response as two dimensions. After the models are applied, they provide various insights into the internal structure of the exam. These insights come in the form of statistical values (see Table 2) at the global level for the entire exam or student population, or at the individual level for a single exam task or student. This is followed by a statistical interpretation that allows inferences to be made regarding exam structure, quality, etc. It should be noted that information on an individual exam task is only meaningful in relation to

TABLE 2

Global statistics for model fit comparison. Itemfit statistics show how well the estimated item difficulty (σ) fits the data.[4,11,12] RMSD is an alternative measure for itemfit.[11,13] Personfit shows how well the model represents the students' responses.[4,14] For these three statistics, mean and standard deviation (sd) are given in the table, summarizing the statistics for every single exam. SRMR is a global fit statistic.[11,15,16] The significant underfit shows the proportion of students with bad representation by the model. It is defined as a personfit value > 1.3.[4] Information Criteria AICc compares the quality of a model with its complexity (number of model parameters).[5,11,17] It is only interpretable between different models for the same data set. Lower values show model improvement. Reliability represents the measurement error [11] Only the main factor reliability (MF) is given for the Rasch testlet bifactor model.

Statistic	SS15	WS16	SS17	WS17	SS18	WS18	SS19
Uni-dimensional Rasch model (1PL)							
Itemfit (mean)	0.99	0.98	0.94	0.97	0.97	0.94	0.94
Itemfit (sd)	0.06	0.12	0.19	0.15	0.14	0.20	0.20
RMSD (mean)	0.03	0.05	0.05	0.05	0.04	0.05	0.04
RMSD (sd)	0.01	0.02	0.02	0.02	0.02	0.02	0.02
SRMR	0.05	0.07	0.06	0.07	0.06	0.07	0.07
Personfit (mean)	0.94	0.87	0.88	0.87	0.88	0.85	0.90
Personfit (sd)	0.26	0.43	0.38	0.36	0.37	0.45	0.38
Sign. underfit	0.04	0.07	0.03	0.06	0.06	0.05	0.06
AICc	6491	5364	9851	7505	11,860	5734	8251
WLE Rel.	0.76	0.88	0.90	0.90	0.91	0.91	0.90
EAP Rel.	0.78	0.91	0.93	0.94	0.93	0.94	0.93
Rasch testlet bifactor model (BF)							
Itemfit (mean)	0.99	0.97	0.94	0.97	0.97	0.95	0.94
Itemfit (sd)	0.05	0.15	0.19	0.14	0.14	0.20	0.19
RMSD (mean)	0.05	0.10	0.08	0.13	0.10	0.09	0.10
RMSD (sd)	0.01	0.02	0.02	0.02	0.02	0.02	0.02
SRMR	0.05	0.05	0.05	0.05	0.05	0.07	0.06
Personfit (mean)	0.97	1.02	0.97	0.98	1.00	0.92	1.02
Personfit (sd)	0.28	0.51	0.43	0.39	0.42	0.49	0.42
Sign. underfit	0.05	0.14	0.07	0.11	0.13	0.08	0.10
AICc	6433	5141	9640	7131	11,318	5625	8008
EAP Rel. (MF)	0.69	0.85	0.90	0.89	0.88	0.90	0.88
Partial credit model (PCM)							
Itemfit (mean)		0.99	0.96	0.99	0.99	0.91	0.96
Itemfit (sd)		0.10	0.16	0.16	0.15	0.22	0.20
RMSD (mean)		0.05	0.05	0.05	0.04	0.05	0.05
RMSD (sd)		0.02	0.02	0.01	0.02	0.02	0.02
SRMR		0.07	0.06	0.07	0.06	0.08	0.08
Personfit (mean)		0.85	0.86	0.87	0.88	0.87	0.90
Personfit (sd)		0.50	0.44	0.40	0.42	0.54	0.44
Sign. underfit		0.04	0.03	0.04	0.06	0.06	0.04
AICc		4796	8883	6408	10,291	5074	7243
WLE Rel.		0.84	0.86	0.87	0.88	0.88	0.88
EAP Rel.		0.88	0.90	0.91	0.91	0.92	0.91

other exam tasks; whether a task is easy or difficult, fair or unfair, rich in information or redundant only makes sense in comparison to the exam as a whole.

The following presents selected results with relevance to our competency model and course development modeling.

Applying the One-Dimensional Model

In a first step we apply the simplest of the three applied models, the uni-dimensional Rasch model. It assumes a single competency to be reflected by the exam results. For this reason, the uni-dimensional Rasch model is also referred to as a one-parameter model. It provides a first insight into data quality and how it responds to the model. The uni-dimensional Rasch model generally fits our data well. This is expressed by several of the statistics in Table 2. First, the personfit shows how well the model represents the empirical data on every student (their estimated level of performance). A value of 1.0 would indicate a perfect fit (the model describes the students' performance perfectly). Usually, values between 0.5 and 1.7 are considered acceptable. The personfit of our Rasch model is grouped closely around 0.9, with an acceptable deviation showing a slight overfit. The itemfit statistics follow the same logic and calculation, applied to the fit of every exam task (item). The model describes the empirical difficulty of the exam very well. The RMSD statistic is an alternative way to calculate the itemfit. Acceptable values are between 0.01 (excellent fit) and 0.08 (acceptable fit). Again, this shows a good representation of the data under the assumption of a Rasch model. Finally, global statistics further confirm a good model fit, with a reliability of 0.75 (barely made it) up to > 0.9 in later semesters (excellent reliability), and a SRMR of < 0.08.

However, the Rasch analysis also indicates that a two-dimensional structure might be applicable. This is indicated by Figure 1, which shows a heat map of the correlation between the residuals of every exam task of the exam (only the exam from summer 2015 is shown as an example). The darker the shading, the more dependency can be assumed between the exam tasks. Each group of exam tasks that are closely related to each other (usually referred to as *testlets* [7] (p. 126) in psychometry) is surrounded by black borders, emphasizing the dependency between exam tasks belonging to the same thermodynamic subject matter. Item correlations between testlets are rather weak. However, there is evidence that some exam tasks where students are expected to apply similar methods or approaches (e.g., the first law of thermodynamics) are also mildly correlated.

This result suggests that it is not just a question of a single competency, but that students tend to use a general and several subject-related sub-competencies to pass the exam. Otherwise, the residuals would correlate rather inconspicuously across all exam tasks. For example, the students' competency might be structured along a general factor (*thermodynamics* as a broad and holistic competency) and subject-specific group factors (e.g., specialized sub-competencies such as the application of the first law of thermodynamics). This is a typical field of application for a Rasch testlet bifactor model, which assumes a general factor and multiple group factors for the students' ability.[19]

Applying the Two-Dimensional Bifactor Model

Our exam was designed following the *constructive alignment* framework we introduced in Part I (see Figure 2) of our series,[1] thence we expected the exam to reflect the learning objectives and the thermodynamic subject matter that define the whole course. Since we structure the exam along five testlets that group several tasks around a common thermodynamic problem (see Figure 2), this could also be represented by the empirical structure of the students' exam responses. We initially assumed students would develop sub-competencies along these different thermodynamic subject matter. However, after fitting the uni-dimensional Rasch and the two-dimensional Rasch testlet bifactor model to the student's responses, the results are ambiguous.

The Rasch testlet bifactor model assumes one general factor and three or four group factors to describe the exam

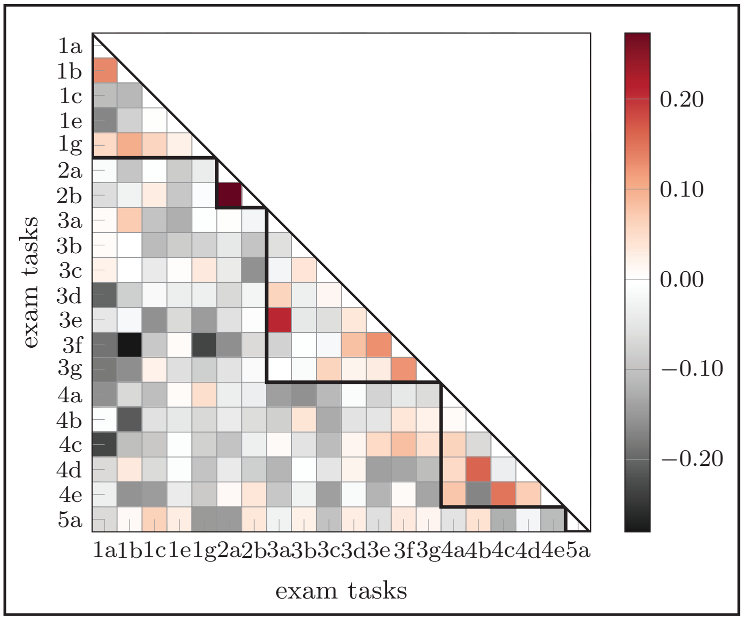

__Figure 1__. Correlation matrix (Q_3 statistic [18]) showing item-correlations within the exam testlets (summer semester 2015). The grouping within each testlet is clearly visible. However, further investigation shows that essential uni-dimensionality of the overall test can be assumed.

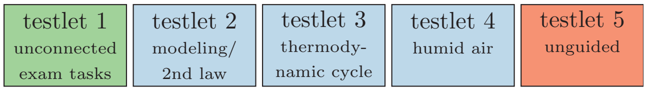

***Figure 2**. Exam structure showing the five testlets (problem sets) and their characteristics (from Figure 5 of Part I [1]). The first testlet consists of independent short exam tasks covering material from all subject matter. It tends to focus on factual knowledge and reproduction of basic facts or concepts. Testlets 2 to 4 are longer guided exercises with several thematically grouped exam tasks, where each exam task can be solved individually. The fifth testlet is a problem without being structured by individual exam tasks. Testlet five is not fixed to a specific subject matter. The color coding intends to show similarity in structure. The final exam alone determines the grade in our one-year course.*

responses, depending on the exact design of the exam in a given year. However, the first testlet is never modeled with group factors; only the general factor is used, since all exam tasks in this testlet are unrelated to each other.

The bifactor model reveals some unexpected results. First, the model quality is inferior to the less complex uni-dimensional Rasch model. This is reflected by the itemfit measured as the root mean square deviation (RMSD), which breaks the 0.08 limit in several cases for the bifactor model. Also, the deviation of the personfit is larger as for the Rasch model, indicating that a significantly larger proportion of the students are not well represented by the bifactor model. This is also expressed by the significant underfit entries in Table 2. These findings imply that the fitted bifactor model offers a worse description of the actual item difficulty and student ability. Additionally, reliability measures are also inferior, with values being 0.03 to 0.06 below the ones for the Rasch model. Thus, we have a high degree of measurement error if we want to look into the students' ability in regard to the single testlets.

More importantly, following Reise, Rodriguez, and Haviland,[19] we tested for uni-dimensionality. This procedure is used to determine whether the role of group factors compared to the general factor is relevant to the model. Table 3 shows the key statistics for the analyzed exams over the past years. As the ratio between the explained variance of the whole model and the general main factor is close to unity, and ECV and PUC are also significantly larger than 0.5, Reise, Rodriguez and Haviland suggest to treat the model as having a strong essential uni-dimensionality. Thus, a strong general factor exists that dominates the group factors in the two-dimensional model.

TABLE 3

Statistics for the assessment of essential uni-dimensionality:[19] ω_t is the measure for reliability in a multi-dimensional model. It represents the proportion of the explained total variance. ω_H represents the hierarchical reliability. It explains which proportion of the variance is explained by only the model's general factor. $\omega_\mathrm{H}/\omega_t$ indicates the proportion of the explained variance caused only by the general factor. ECV is the *explained common variation* and expresses the relative strength of the general factor. PUC is the *percentage of uncontaminated correlations*. It puts the number of item correlation between testlets in relation to the number of item correlations *within* testlets.

Semester	$\omega_\mathrm{H}/\omega_t$	ECV	PUC
Summer 2015	0.875	0.630	0.830
Winter 2016/17	0.913	0.663	0.817
Summer 2017	0.964	0.847	0.798
Winter 2017/18	0.930	0.745	0.812
Summer 2018	0.927	0.729	0.815
Winter 2018/19	0.968	0.886	0.785
Summer 2019	0.930	0.749	0.802

Essential uni-dimensionality might indicate that the exam is not so much about *dealing with distinct thermodynamic subject matter* as it is about undesired *dealing with the test*. This suspicion is further supported by results from qualitative data shown in the results section of Part I.[1] Students prepare for the exam primarily through repetitive practice using old exams. Many students do not necessarily study towards mastering subject-related concepts and challenges, but rather specialize in pragmatically coping with the exam. This is discussed in educational research as *test wiseness* or, less euphemistically as *surface level of learning*.[20] It implies that students actively prepare themselves for the exam situation and the unique, expected exam design. They then demonstrate a certain competency in dealing with the exam situation that may supersede, or at least rival, the conceptual understanding, declarative knowledge, and procedural abilities they show in regards to the different thermodynamic subject matter. Essential uni-dimensionality could also imply that our exam design or the students' way of learning conceals an actual competency-structure beneath the surface of our observable data. This can be caused by the time constraints of the exam situation, but also by the number and the scope of the exam tasks. The exam in its current form would then not be able to reveal this structure, and competency-dimensions might not get uncovered for the different subject matter of thermodynamics.

On the other hand, one can draw a more positive conclusion. The indication of essential uni-dimensionality also means that our exam measures competency in a holistic approach. Regardless of the subject matter, procedures, and types of knowledge actually acquired by the students, the exam allows them to engage in a complex and realistic challenge. The exam tasks we confront them with are based on general thermodynamic problems drawn from major subject matter of the discipline. Therefore, uni-dimensionality has a positive connotation as well. We do not measure fractured knowledge, as would be the case with an exam where students specialize in one subject matter at the expense of another. On the contrary, it is highly desirable (as we outlined in Part I [1] and above) that the students achieve an understanding of a strategy of general problem solving in engineering thermodynamics (*subject-specific problem solving*), independent of specific applications. It should also be noted that essential uni-dimensionality has positive methodological implications for exam modeling and further analysis of the results. A methodological advantage of uni-dimensionality lies in the fact that it is possible and justified to use less complex and more robust psychometric models in order to represent the exam data. For example, the uni-dimensional Rasch model provides statistics for item difficulty and person ability with better reliability (less measurement error) and global fit statistics than the Rasch testlet bifactor model. If we have more robust statistics for student performance, we are able to evaluate our test fairness and the quality of the final grades. A well-fitting statistical model therefore makes the evaluation of the exam quality much more reliable.

Using the uni-dimensional Rasch model and the Rasch testlet bifactor model, we were able to provide a quality measure of the students' performance in the final exam, but we were only able to measure a uni-dimensional holistic thermodynamic competency that lacks essential insights into different dimensions of competency applied by the students. We have to inquire further to reveal the empirical competencies our students actually acquire throughout our course. We continue by shifting the perspective from dimensions of competency *between* exam tasks to one that focuses on different dimensions *within* single exam tasks.

Applying the Partial Credit Model

Based on the results from the bifactor analysis and the observation of essential uni-dimensionality, we conclude that our exam does not primarily represent a competency-structure by thermodynamic subjects *between* individual tasks or testlets. Instead, we shift our perspective to dimensions *within* individual exam tasks and find *thermodynamic modeling* and *mathematical solution* as two overarching dimensions of competency.

A partial credit model assumes that different kinds of competencies or levels of achievement are necessary in order to solve a single exam task. In order to apply such a model, changes to the way the exam was designed and graded became necessary. As a result, we made significant changes to the design and assessment procedures in the following years. We introduced a two-stage grading process where graders now grade the correct *modeling approach* (AP) and the correct *quantitative solution* (SP) separately, as two distinct aspects (as described in the results section of Part I).[1] This grading approach aligns well with *subject-specific problem solving* proposed in Table 4 of Part I [1] and is a major step towards a grading system that goes beyond simple summation of scattered points without regard for qualitative differences among student performances. Ideally, an exam task is now graded according to three achievement levels: incorrect modeling approach and no actual quantitative solution (0), correct modeling approach but missing or wrong quantitative solution (1), and correct modeling approach and correct quantitative solution (2). Some exam tasks do not require a quantitative solution. This is easily integrated into the modeling process.

In conclusion, we observe that the partial credit model achieves the best model fit for our exam data, allowing for a robust measure of student performance and exam task difficulty. It supersedes both the uni-dimensional Rasch and Rasch testlet bifactor model in regard to how well it describes the actual exam data. Further details of the model comparison will be discussed in the following section. See Table 2 for details and references.

Comparing the Results and Quality of All Three Models

The statistical analysis reveals that all three models show largely acceptable itemfit measures, which indicate how well the models fit the data in regard to the difficulty parameters σ, estimated for every single exam task in an exam. A good fit is considered to range between 0.75 and 1.3. The σ-parameters are the core estimates for a model. They describe the relative difficulty of the individual tasks in our exams. The values in Table 2 show that all three models are acceptable in order to describe the basic structure of our exam in regard to item difficulty σ. The RMSD (root mean square of deviation), as a non-central alternative itemfit statistic, basically confirms this finding. Values of RMSD below 0.05 are considered a very good itemfit. However, we see that the bifactor model clearly has a higher RMSD. Since the RMSD is a non-centrality parameter, we interpret this observation as a lack of test power in the bifactor model. In regard to the global fit statistic SRMR that represents the overall model fit in contrast to the ones on single item or person level, we see that all three models perform well. Values below 0.08 are considered a good fit, with 0 being the total identity between observed and modeled distribution.

Also, as shown by the personfit means, standard deviations, and the proportion of significant underfits, all three models produce largely comparable personfit statistics. They, too, should fall between 0.75 and 1.3. However, the partial credit model shows the best result in regard to significant underfit of the person ability estimators θ. If a student has a significant underfit (personfit value > 1.3), the model fails to properly describe this student's exam performance. The partial credit model slightly outperforms even the otherwise very robust uni-dimensional Rasch model. It represents the students' performance with highest precision and underlines that thermodynamic modeling and quantitative solution are two distinct competencies. This is of importance because we are interested in the assessment of students, their competency-structure, and the quality of the exam (e.g., test fairness).

If we turn to the information criterion AICc, we see the strongest indication for the superiority of the partial credit model. The AICc is a measure for the amount of information a model provides for a given set of data (a single exam). It compares the likelihood of a model to its number of estimated parameters. An increase of parameters is penalized. The AICc is only meaningful when compared between different models for the same data set (exam). Lower values indicate a better model quality and efficiency. Even though the partial credit model introduces a significant number of new parameters (the threshold parameters for the partial credits), the information criterion improves greatly.

We summarize that the partial credit model equals the robust uni-dimensional Rasch model in regard to itemfit and global fit. It is slightly superior in regard to personfit and greatly improves the relative model quality, as measured by the information criterion. Modeling the dimensions of competency *within* single exam tasks seems to be much more fruitful than our previous search *between* items and testlets. The incorporation of the two distinct student performances (finding the right modeling approach to a problem and the performance of its quantitative solution) into the model was a big step forward. We consider this an important result for properly defining dimensions of competency. The following examples illustrate the significance of this result as they allow for a deeper insight in how task difficulty is created and how different requirements are hidden within a single task.

In Figure 3, Thurstonian thresholds [10,11] indicate the relative difficulty of exam tasks that require the determination of the thermodynamic efficiency of a thermodynamic cycle (testlet 3). The exam task is very similar in all exams, but it is never exactly the same. The partial credit model deciphers the subtle differences. The solid lines show the difficulty of a single exam task for its correct modeling approach. By convention the difficulty is compared on the p = 0.5 level. Thus, if a graph passes this level at θ = -1.2, it is considered less difficult than another graph that passes the threshold at θ = -0.5. It can be seen that the exam tasks on finding the right modeling approach for calculating thermodynamic efficiency are very similar in difficulty over the considered years. All the solid lines are located within roughly a one θ *logit* span. The modeling approach for this kind of exam task is a good candidate for a possible anchor-item for linking different exams over time. However, the quantitative solutions, indicated by the dashed lines, show substantial changes in difficulty over time.

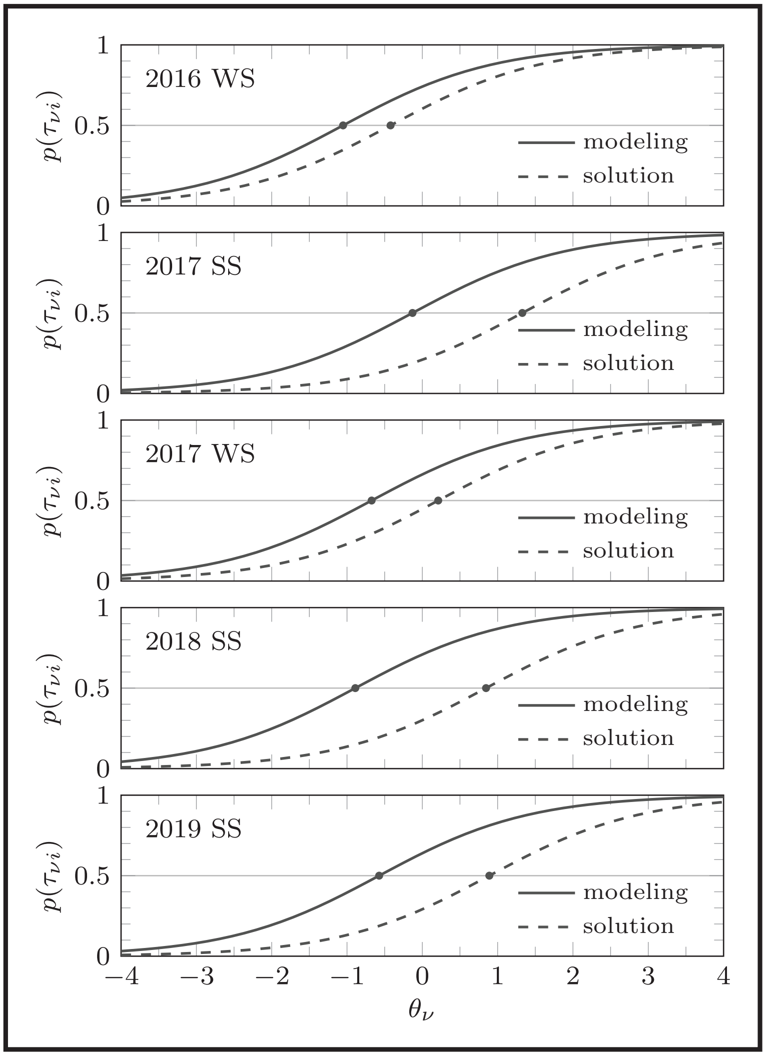

***Figure 3**. Thurstonian threshold curves $\tau_{\nu i}$ for the partial credit model differentiating between modeling approach (solid lines) and actual quantitative solution (dashed lines) for the calculation of the thermodynamic efficiency of a thermodynamic cycle (testlet 3) for different exams between the winter semester (WS) 2016 and the summer semester (SS) 2019. Depicted is the probability p of person ν with ability θ_ν for achieving partial credits for several exam tasks i. The winter exam 2018 did not contain an exam task for thermodynamic efficiency.*

In 2016, finding the correct modeling approach and correctly solving the mathematical calculations was only a minor step in difficulty. One semester later, the difference was several times larger. The mathematical challenge had increased significantly. This kind of data enables us to better understand what actually makes exam items easy or difficult.

The second example in Figure 4 adds a different interpretation. Two items are compared in which both ask students to determine the mass flow rate in a humid air exercise (testlet 4). The difference in overall difficulty and the distance between difficulty for modeling approach and quantitative solution can be caused by several things that differed notably between the exams:

- The task description of the more difficult exam task (2017) is more complex.
- Different context of the problem (air conditioning vs. air purification).
- The exam task from 2017 requires the determination of enthalpy as an intermediate step and thus requires a longer quantitative solution.

The partial credit model helps to identify different factors that determine the difficulty of an exam task. It also shows that similar exam tasks are heavily influenced by context factors and, more importantly, that modeling approach and quantitative solution are two distinct dimensions of competency. As a result, the partial credit model provided us with a statistically robust tool to explore two distinct dimensions of competency, thermodynamic modeling and quantitative solution, as well as a way to investigate other influences on the difficulty and structure of the examination. We build on this in Figure 6.

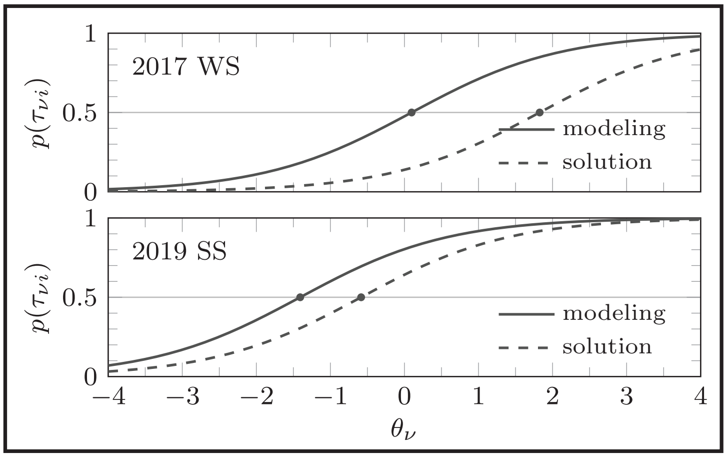

***Figure 4**. Probability p of person ν with ability θ_ν for achieving partial credits for several exam tasks i. Two items on mass flow rate in humid air from different exams are compared. The graphs show Thurstonian thresholds $\tau_{\nu i}$ for modeling approach (solid lines) and actual quantitative solution (dashed lines). Considered are the exam in the winter semester (WS) 2017 and the summer semester (SS) 2019.*

PRACTICAL APPLICATIONS AND RESULTS

What is in it for the students? Based on the results from the previous section, we report how we improved our exam quality over time in terms of test fairness. We also analyze what makes exam tasks easy or difficult in order to gain a deeper insight in the empirical challenges students experience in our course on thermodynamics.

Improvement of Exam Quality

The analysis of our exams provides us with decisive insight into possible underlying dimensions of competency. Far beyond this benefit, it has also significantly improved – and continues to improve – the quality of our exams in general. Generally speaking, with the introduction of exam modeling, we started to treat our exam as a psychometric test. This implies that we apply quality measures such as objectivity, validity, and reliability to the exam design process to gain deeper insight into test quality [21,22] (pp. 27-43 and p. 15, respectively). As a result, quality guidelines were established [23] and are now used throughout the exam design process (see Table 4). They help the changing team of teaching assistants to keep fundamental quality requirements in mind when designing and reviewing exam tasks. (The professors sign off on the exams after several rounds of feedback between them and the teaching assistants.) Table 2 documents the continuing improvement, e.g., by a continuous increase of test reliability (WLE and EAP reliability) over time. This basically means that the measurement error of our exam has decreased and the exam became a more and more reliable indicator of student performance.

Exam modeling also had a major impact on the quality of the grading process. The used cumulative grading system determines the final grade by summing up all grading points across the different testlets and exam tasks. Even though we attempted to distribute grading points according to different levels of difficulty, this is a blurry and imprecise way of determining the final grades of the students. It is, of course, intuitive and readily accepted by the students. The fitted partial credit model, as an estimate for the students' latent person ability θ is more robust and accurate because it takes into account that every exam task has a unique level of difficulty. The modeling process reveals the actual difficulty of exam tasks with more precision and reliability than we could have achieved with a personal estimate and goodwill judgment alone.

TABLE 4
Quality guidelines for the design of the final exam.

Exam tasks are aligned with the learning objectives and the covered subject matter and are tested only with familiar methods.

The style of the exam is transparent for the students; no new type of problems or concepts are introduced in the exam.

Exam task descriptions and problem statements are short, clear, familiar, and should not pose a language barrier; omit non-essential information, and require no specific social or professional background, i.e., the exam tasks are formulated as thermodynamic problems not necessarily framed in a larger context (no need for abstraction).

Difficulty of the exam tasks increases gradually and continuously to better discriminate different levels of competency; difficult exam tasks are located towards the end; smaller tasks have proven to be easier.

The exam can be passed by solving the exam tasks that define the minimum level of competency required in the course.

Each exam task can be solved individually without dependence on previous ones, necessary intermediate results are provided (solving an unguided thermodynamic problem is tested in testlet 5); general and specific tasks for the same topic do not exist.

The exam can be completely solved in the given time frame to reduce the effect of test wiseness/dealing with the test; difficulties are introduced by difficult exam tasks, not by introducing a time constraint.

Different modeling approaches (e.g., for physical properties) are tested in each testlet; avoid duplicates.

Subject-specific problem solving is used as a guideline for composing and grading exam tasks; this is explicitly known to the students.

The exam is proof-calculated by a colleague uninvolved in the exam design; not only to find potential errors, but also to reduce personal bias.

The exam is created as early as possible to allow didactic consequences to be incorporated into the course.

Grading guidelines, e.g., based on a solution manual, are agreed upon among all graders and are known to the students; thermodynamic modeling and quantitative solution are graded as two different competencies.

Over the past years, we observed the difference between the estimated ability-level of students based on traditional cumulative scores and on the partial credit model. We found that the differences steadily decreased as we improved the design process of the exam.

Figure 5 illustrates this development by comparing the estimated person abilities θ against the traditionally determined cumulative scores for six past exams. Ideally, all cases would fall on a single curve without scattering. From top to bottom (i.e., in chronological order), a constant improvement towards this goal is visible. This implies an improvement of test fairness since fewer students suffer from inaccuracies inherent to cumulative grading. Because legal concerns prohibit the use of probabilistic test theory as a grading method, we continue to improve the exams in this way and further reduce inaccuracies. The partial credit model provides us with the means to assess this process.

A major step in this improvement in exam quality was the introduction of the two-level grading system proposed in combination with the partial credit model that documents the difference between the students' abilities in thermodynamic modeling and actual quantitative solution of thermodynamic problems. As a result, the partial credit model has enabled us to gain deeper insight into the underlying dimensions of the students' abilities documented in the exams. As a future perspective, this is an important first step in further improving the grading process by taking into account the actual empirical difficulties and different competencies students apply in the exam.

What Makes Exam Tasks Easy or Difficult?

The results from the partial credit model enable us to estimate the empirical difficulty of different exam tasks (by ordering the Thurstonian threshold curves as, for example, depicted in Figures 3 and 4). This allows us to identify easy and difficult exam tasks and what makes exam tasks easy or difficult for students, which facilitates awareness about general difficulty traits. We approach the analysis of the difficulty of exam tasks as follows. First, we arrange all exam tasks of the final exam in terms of their empirical difficulty.

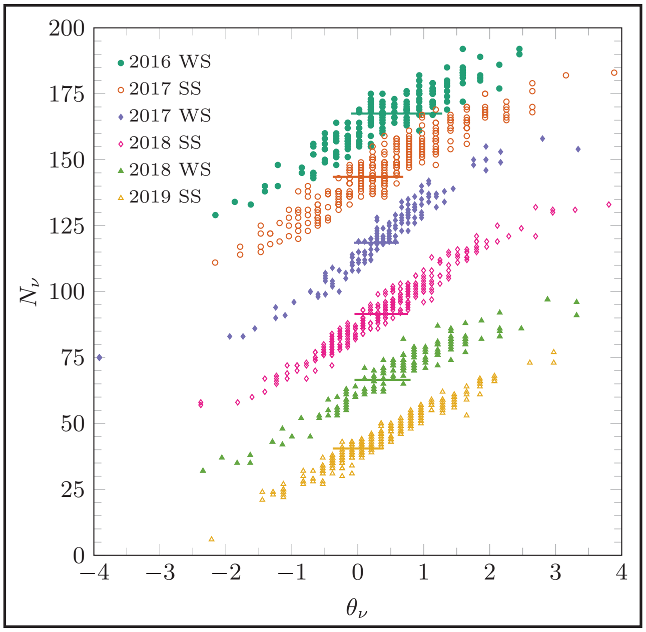

Figure 5*. Comparison of the cumulative evaluation results N (total number of points) to the person ability θ of all exam participants ν and different exams (different exams are shifted vertically by 25). The horizontal lines represent the minimum number of points in the summative evaluation required to pass the exam.*

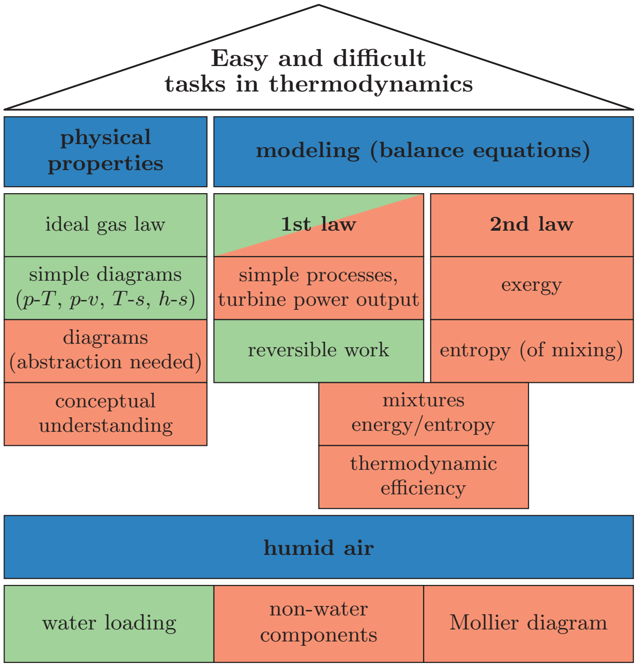

Figure 6*. Visualization of subject matter of easy (green) and difficult (orange) exam tasks from the analysis of final exams between the winter semester 2016 and the summer semester 2019, classified into the categories physical properties, modeling (balance equations), and humid air (blue).*

Second, we then analyze across different exams which traits are frequently more difficult than others. The classification of traits could be testlet number, topics from the subject matter (see Figure 4 of Part I), points for the thermodynamic modeling approach and quantitative solution, and our estimate of required mathematical skills. Third, we look for consistently difficult or easy traits, not only by an absolute measure, but also by looking at outliers and (to us) surprisingly easy or difficult exam tasks.

Reliable difficulty traits are the topics from the different subject matter. The result of our analysis is summarized in Figure 6. We classified the easy and difficult traits in three main categories: physical properties, balance equations, and humid air. While exam tasks covering the ideal gas law or simple diagrams were consistently easy for the students, the opposite is true for physical property exam tasks that include more complex diagrams and a requirement for good conceptual understanding. This might be due to the different approaches to learning: *deep* versus *surface* level of understanding. Balance equations (first and second law of thermodynamics) appear to be among the difficult subject matters in most cases. While the first law and energy calculations may be easy in one context, especially for calculating reversible work, it can be difficult in another context. The second law, including exergy balances and entropy calculations (especially when including mixtures), seems consistently difficult for students. Although it is a recurrent exam task, calculation of the thermodynamic efficiency of thermodynamic cycles seems difficult for the students as well. Exam tasks that fall in the humid air category and are easy for the students to deal with are those concerned with the calculation of water loading, while treatment of substances other than water in the mixtures as well as the use of Mollier diagrams challenge the students.

One point that we have noticed throughout is that long problem statements as well as non-essential information make the exam tasks needlessly more difficult, compared to exam tasks where students are required to perform comparable tasks.

CONCLUSION AND OUTLOOK

With this work we provide an example of what can be learned from applying probabilistic test theory to an exam in an engineering thermodynamics course to assess student performance and empirical competencies. We are at the beginning of a deeper understanding of competency levels based on probabilistic test theory. Proper statistical procedures are still in the state of development and evaluation.[24] However, our results show that a significant difference between assumed and empirical competencies can be observed.

Conclusion

We analyzed six final exams of our engineering thermodynamics course and fitted three item response theory models with respect to their capability as performance measures. The partial credit model shows the best correlation-quality of all three models and can be recommended to measure student performance. The statistical analysis of this work adds parts of significant importance to a thermodynamic competency model, as the final exam is the most important, single representation of student performance throughout the course.

The available results are also valuable for analyzing the dimensions of competency in our students' performance. We now have a statistical tool that allows us to make meaningful and solid interpretations of the qualitative findings on *student learning* and their difficulties. With regard to *constructive alignment*, we found that the empirical dimensions of competency of our exam, which is strongly built around thermodynamic subject matter, are *thermodynamic modeling* and *quantitative solution* as supported by the partial credit model. This aligns well with the competencies we want our students to acquire, if the quality guidelines for exam design from Table 4 are followed.

We want our students to acquire a conceptual understanding of thermodynamic principles that they are able to apply across different subject matter. While the tested competencies align well with the competencies we want our students to acquire (*learning objectives*), the students' learning preferences towards *exam proficiency* [1] (*learning activities*) indicate misalignment thereof and the exam design. We responded by making it transparent to the students that the application of thermodynamic problem-solving strategies and the subsequent use of mathematical methods is the core aspect of what we expect from them (concentrated as *subject-specific problem solving*, proposed in Part I, Table 4 [11]). We also heavily emphasized this point towards the students in our lectures, plenary exercises, and guided tutorials. The new two-level grading procedure has done much to improve transparency and understanding of the competency we wish to facilitate.

Furthermore, we were able to improve the test fairness of our exam by following the quality guidelines in Table 4.

Outlook

Looking forward, we aim to implement further measures to improve our engineering thermodynamics course. For example, new types of exam tasks could be introduced in order to test for certain aspects of competency more specifically. The revision of the exercises with respect to clear and simple task descriptions to match the final exam and *cognitive apprenticeship* in the solution manual used in the plenary exercises and the guided tutorials has not yet been completed. As mentioned above, the partial credit model allows for a more elaborate grading process that overcomes some weaknesses of the traditional (summative) grading approach, in particular by accounting for the different competencies students apply in the exam and the empirical difficulties of exam tasks that may unexpectedly differ from intuition during the exam design process.

Another contemplated measure is the separation of the final exam into two exams, one after each semester, to lower the barrier of just one large exam that alone determines the final grade of the course. This would require a change in the course structure, which would have a direct impact on the learning objectives and thus on the competency-structure, which would then have to be reassessed. Possible changes could include focusing on thermodynamic concepts in the first semester, while the second semester focuses on more demanding thermodynamic cycles with technical relevance. The purpose of this approach would be to shift the students' perception away from the mere coping with the exam (which focuses on technically relevant thermodynamic cycles) towards a deeper understanding of thermodynamics fundamentals, as already envisioned in the learning objectives. This would, of course, require changes in the study material, guided plenary exercises and the guided tutorials as well as the qualification of the student teaching assistants.

ACKNOWLEDGMENTS

The authors thank the German Federal Ministry of Education and Research (BMBF) for financial support through the project *Qualitätspakt Lehre – Individualität und Kooperation im Stuttgarter Studium (QuaLIKiSS)* and the Foundation Innovation in Higher Education for financial support through the project *digit@L: Digitales Lehren und Lernen an der Universität Stuttgart – Boost. Skills. Support*. We thank the German Research Foundation (DFG) for supporting this work by funding – EXC2075 – 390740016 under Germany's Excellence Strategy. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). We especially thank Irene Kreitmeier and Sonja Rapp for initiating the continuous improvement of the course *Engineering Thermodynamics 1/2* as part of QuaLIKiSS, and Niels Hansen for his ongoing guidance of the continuous improvement of the examinations and for his help improving this manuscript. We thank our colleagues Gernot Bauer, Jörg Baz, Florian Drunsel, Johannes Eller, Timm Esper, Maximilian Fleck, Julia Gebhardt, Andrea Hemmen, Madlen Hopp, Christopher Keßler, Timo Klenk, Christoph Klink, Hamzeh Kraus, Alexander Lange, Oliver Lötgering-Lin, Jonas Mairhofer, Daniel Markthaler, Philipp Rehner, Anja Reimer, Elmar Sauer, Marina Stavrou, Marc Theiss, Christian Waibel, Dominik Weidler, Julia Zmpitas and Wasilios Zmpitas for their help with the not always easy implementation of the insights gained in this endeavor.

REFERENCES

1. Stierle R, Fischer M, Braun T, and Gross J (accepted for publication) Investigating learning and improving teaching in engineering thermodynamics guided by constructive alignment and competency modeling: Part I. Improving our learning environment – How we support student learning. *Chem. Eng. Ed.* 57(2) DOI: 10.18260/2-1-370.660-126287
2. Braun T (2019) Evaluation Research for Study Success Requirements ITT 2014/15. DaRUS. DOI: 10.18419/darus-470.
3. Braun T (2018) Klausurmodellierung in der Technischen Thermodynamik mittels Item Response Theory: Ergebnisse einer hochschuldidaktischen Begleitforschung. *Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF) 2018*. Basel, Switzerland. DOI: 10.18419/opus-10390
4. Boone WJ, Staver JR, and Yale MS (2014) *Rasch analysis in the human sciences*. Springer Dordrecht. Dordrecht, Netherlands. DOI: 10.1007/978-94-007-6857-4
5. Rost J (2004) *Lehrbuch Testtheorie – Testkonstruktion,* 2nd edition. Huber, Bern, Switzerland.
6. Wainer H, Bradlow ET, and Wang X (2007) *Testlet Response Theory and Its Applications*. Cambridge University Press. Cambridge, England. DOI: 10.1017/CBO9780511618765
7. Wang WC and Wilson M (2005) The Rasch testlet model. *Applied Psychological Measurement* 29(2):126-149. DOI: 10.1177/0146621604271053
8. Geiser C and Eid M (2010) Ch 4: Item-Response-Theorie. Wolf C and Best H. *Handbuch der sozialwissenschaftlichen Datenanalyse*. VS Verlag für Sozialwissenschaften. Wiesbaden, Germany. DOI: 10.1007/978-3-531-92038-2_14
9. Thissen D and Steinberg L (1988) Data analysis using item response theory. *Psychological Bulletin* 104(3):385-395. DOI: 10.1037/0033-2909.104.3.385
10. Masters GN (1982) A Rasch model for partial credit scoring. *Psychometrika* 47:149-174. DOI: 10.1007/BF02296272
11. Robitzsch A (2016) Test Analysis Modules (R package 'TAM'). https://cran.r-project.org/web/packages/TAM/TAM.pdf.Accessed 14.01.2017.
12. Wright BD and Masters GN (1990) Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions* 3(4):84-85. URL: http://www.rasch.org/rmt/rmt34e.htm
13. MacCallum RC, Browne MW, and Sugawara HM (1996) Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods* 1(2):130-149. DOI: 10.1037/1082-989X.1.2.130
14. Robitzsch A. (2016) Supplementary Item Response Theory Models (R package 'sirt'). https://cran.r-project.org/web/packages/sirt/sirt.pdf. Accessed 15.01.2017.
15. Maydeu-Olivares A (2013) Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives* 11(3)71-101. DOI: 10.1080/15366367.2013.831680
16. Kenny DA (2015) Measuring Model Fit. http://davidakenny.net/cm/fit.htm. Accessed 29.12.2016.
17. Bühner M (2011) *Einführung in die Test- und Fragebogenkonstruktion*. 3rd edition. Pearson Studium. München, Germany. URL: https://elibrary.pearson.de/book/99.150005/9783863268138
18. Yen WM (1984) Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement* 8(2):125-145. DOI: 10.1177/014662168400800201
19. Reise SP, Rodriguez A, and Haviland MG (2016) Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods* 21(2):137-150. DOI: 10.1037/met0000045
20. Biggs J and Tang C (2011) *Teaching for Quality Learning at University: What the Student Does*. 4th edition. Open University Press. Maidenhead, England.
21. Schaper N and Hilkenmeier F (2013) Umsetzungshilfen für kompetenzorientiertes Prüfen: HRK-Zusatzgutachten. *Gutachten der Hochschulrektorenkonferenz*. URL: https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-03-Material/zusatzgutachten.pdf
22. Arbeitsstelle für Hochschuldidaktik der Universität Zürich (2017) Leistungsnachweise in modularisierten Studiengängen. *Dossier Unididaktik*.
23. Braun T (2018) Die Klausur als Orakel? Arbeitsergebnisse einer Klausurmodellierung in der Technischen Thermodynamik. *zlw Working Paper 5*. DOI: 10.18419/opus-9858.
24. Leuders T and Sodian B (2013) Inwiefern sind Kompetenzmodelle dazu geeignet kognitive Prozesse von Lernenden zu beschreiben? *Zeitschrift für Erziehungswissenschaft* 16:27-34. DOI: 10.1007/s11618-013-0381-5
25. Society of Manufacturing Engineers (2012) *Competency Model: Certified Manufacturing Technologist (CMfgT) and Certified Manufacturing Engineer (CMfgE)*. Society of Manufacturing Engineers. Southfield, Michigan https://www.sme.org/globalassets/sme.org/training/certifications/technical-certification/competency-model.pdf
26. Ştefănică F, Behrendt S, Dammann E, Nickolaus R, and Heinze A (2015) Ch 5: Theoretical Modelling of Selected Engineering Competencies. Musekamp F and Spöttl G. *Kompetenz im Studium und in der Arbeitswelt – Competence in Higher Education and the Working Environment, Vocational Education and Training: Research and Practice*. Peter Lang. Bern, Switzerland. DOI: 10.3726/978-3-653-04168-2❒