Validation Approach
For Occupational Standards

Institute for Technology, Work Processes and Vocational Education (iTAB)
at the Hamburg University of Technology (TUHH)

by
Henning Klaffke

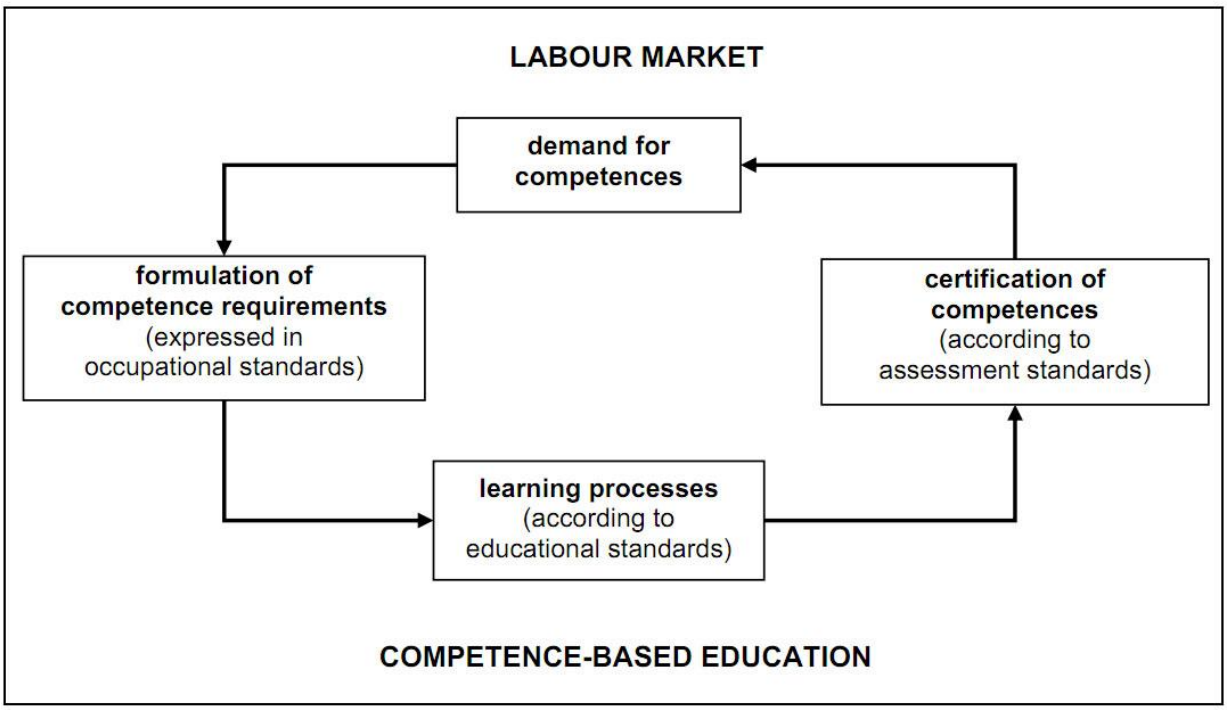
2012

Content

1 Intro of Validation

The concept of validation for occupational Standards is a procedure developed during a scientific project to ensure quality of occupational standards.

Figure 1: **The feedback-loop between the labour market and education**



Referring to Gielen (2000) and Cedefop 2009 the feedback loop between the labour market and vocational education is a continuous process of development of occupational standards, educational standards and assessment standards to ensure high qualified workforce for the labour market.
The development and afterwards the validation of occupational standards is an important step in order to create suitable educational standards and assessment standards.

2 Process of Validation

The methods of validation of occupational standards is a mix of vocational scientific methods like "Experten-Facharbeiter-Workshops", functional mapping and DACUM and classical methods of test development used in psychology. Core of the validation approach for occupational/competency standards in this document are two-step validation connected to a field test. During the development and analysis of this field test two iterations are conducted to empower and strengthen the quality of occupational standards.

2.1 Framework of the validation approach

Figure 1 on the next page illustrates the general scheme of Field test development. The development of tests is realized through a linear sequence of working steps, which will be further described in the

following document. In general, the process is straightforward. However, the working step 2 initiates the first Validation. During content analyzing of the standard and test item generation of the field test the first improvement of the standard is succeeded. Moreover, the working step labeled "Field Test Analysis" comprises activities, like the generation of new test content and crosschecking with the occupational standard. The 2 Validation is based on the outcome of the statistical analysis and the test item description.

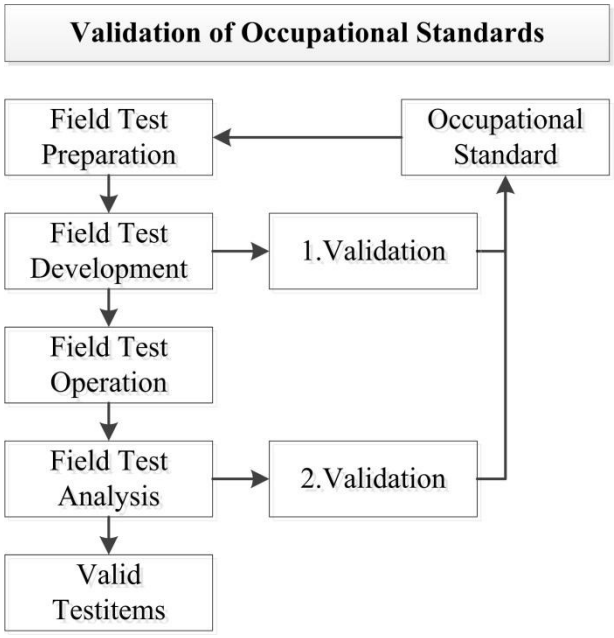


Figure 1: Framework of the Validation Approach

2.2 Purpose and Dimensions of Field Test

The Field Test has two purposes. The first purpose is the strengthening of the quality of the occupational standard. The second purpose is to identify good or bad performing test items in order to produce valid Test items for future assessment. The specification and test dimensions are connected and directed linked to an occupational standard. For the purpose of field test development, a two-dimensional test item structure seems to be appropriate. Given that test scores shall classify test takers regarding minimum competence for the full spectrum of work activities in an occupational field, items must refer on the content level to the identified Major competencies of the occupational standard. Thus,

Major competencies – as stated in the occupational standards – are the primary facet or dimension of the test structure. Content validity of the test requires the inclusion of items for all Major competencies of an occupational standard.

Furthermore, because major competencies are not in a rank order, items have to be in an even balance, e.g. regarding their number and difficulty. If items across major competencies are

unbalanced, it is very likely that test scores show a bias towards those major competencies that have more items in the test. Consequently, performance on these items will have a greater influence on the test scores (positively or negatively). An unacceptable distortion of test results also occurs if the items related to a work process have no balance in their difficulty (as a group too easy or too difficult compared to other item groups).

A second structural dimension (facet) of the field test is the distinction between items that require either an active or a receptive performance from test takers. For example, active performance rests on cognitive activities like analyzing, evaluating, and creating. Obviously, test-items that confront the test taker with problem solving content will prompt such activities. Receptive or reproductive performance, on the other side, is associated with activities such as remembering, understanding, and applying. With other words: the test taker has to use an acquired schema in order to solve the item; usually it is not possible to reveal the schema by deep thinking and deductive reasoning.

The introduced distinction corresponds with the well-known and widely applied taxonomy originally provided by Bloom (1956) and recently revised by Anderson & Krathwohl (2001). Instead of using it in full detail, the six categories of Bloom's taxonomy are compressed, for pragmatic reasons, into the two broad dimensions. The main rationale for using active vs. receptive performance as a second facet/dimension of test structure is the notion that this is an important classification aspect of working tasks. From everyday observation can be inferred that entry-level workers in an oocupational sector (e.g. graduates from vocational training / hospitality institutes) should possess relatively more competencies regarding receptive performances than active performances; nevertheless, it is desirable that expert workers must have the potential of being active members of the work organization. This in mind, the final test shall contain items requiring (prompting) active and receptive performance in a ratio of 2:3.

2.3 Test Difficulty Range

Second, there is a difference between novice and experts. Experts in a domain are highly efficient workers who have acquired (mainly by informal and experiential learning) competencies (knowledge, skills, and abilities) required for successful performance under conditions of time pressure, high stakes, ambiguity, and uncertainty (e.g. Zsambok & Klein, 1997).

Novices are beginners that have had no or very limited experience of the situations in which they are expected to perform. By means of guided participation, they acquire competencies and move on in their career (hopefully) to full participation as an expert in an organization or community of practice (Lave & Wenger, 1991).

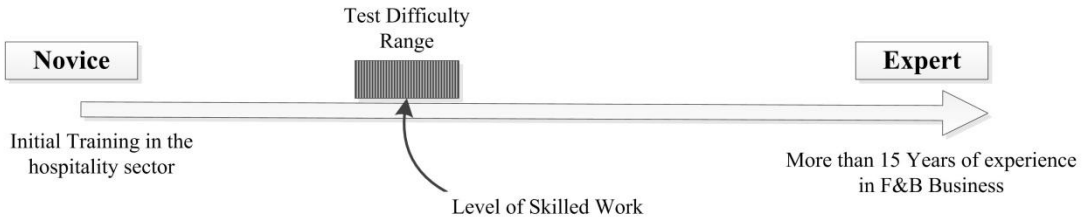


Figure 2 Test Difficulty Range

By referring to a test difficulty range (3-4 years of work experience), we demand that test scores can classify advanced vocational beginners who have coped substantially with real situations (for example during apprenticeship, initial training, further education) and are ready to apply for a first employment. Figure 2 provides a schematic view on this test purpose. As visualized in the figure, the test difficulty range (virtually) span a distance right (difficult items) and left (easy items) from the virtually entry-level point.[1]

2.4 Organising Field Test

Four-option multiple-choice (one correct and three incorrect answers or distractors) has been chosen as general test item type for a field test. In addition to inaugurate a first test it is helpful to allow only question format and avoid completion format in the construction of item stems.

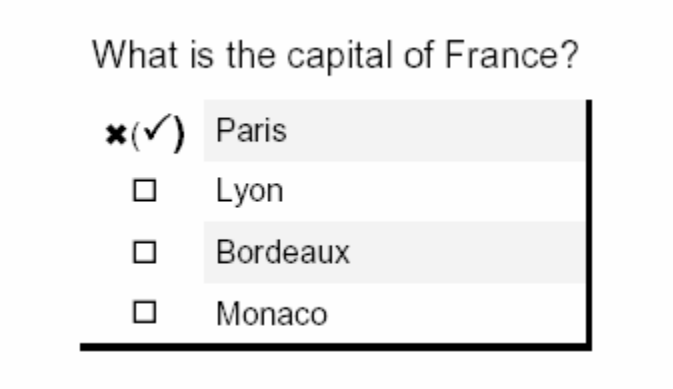


Figure 3 Test Item Characteristic

Figure 4 gives a simple example of a multiple-choice item from the field of Geography. At the top, the test taker will find a question or problem statement (called the item "stem"). In this example, the stem is "What is the capital of France?" Item stems can include additional information like pictures and even movies (if the presentation of the item makes use of the computer). Below the item stem, the test taker finds four answering options. Although all options are plausible answers to the question (well known cities in France in the given example), one option clearly is the right answer (=key) whereas the other three options (=distractors) are clearly wrong. The correct answer in the

[1] The entry-level point needs identification. In the test development schema, this is realized via sampling in the step of item field test. Alternatively or in combination, entry- level competence can be specified by means of expert judgments.

given example is “Paris”. A test taker who knows the correct answer will mark this option and thereby demonstrate his/her competence in Geography.

The test literature offers an extensive list of item construction rules (e.g. Haladyna & Downing, 1989). For example, items generally should minimize test taker reading time, focus on a single problem, and use vocabulary that is consistent with test taker level of understanding. In addition, stems should contain most positive phrasing and avoid negative phrasing. Finally, options should not overlap, be consistent in length, avoid words like “never” and “always”, and be homogeneous in content.

Multiple-choice items have advantages and disadvantages. A clarification is important in order to have a clear understanding of testing limits. First, there is a variety of arguments towards the use of multiple-choice items in the context of testing competence.

- The item format is well known and self-explaining. Thus, test takers – for example, with different cultural or educational background – often understand without instruction what they are supposed to do in order to carry out the test. Familiarization with this item format also facilitates the participation of practitioners in the process of item generation, which is of special importance in field test development.
- The production of multiple-choice items requires relatively little effort. Bearing in mind that a number of items inevitably are objects to replacement or revision after the field test, this fact make multiple-choice items very economical compared to alternative item formats. The relative advantage of multiple-choice items is also clear with regard to practitioner participation in test development: Although the availability of these people is limited, it is possible to generate a number of items in a workshop session that are more or less ready for field test. In addition, the delivery (in a paper & pencil or computer-based format) of multiple-choice items and the administration of multiple-choice based tests needs only moderate resources.
- With multiple-choice items, it is possible to cover a broad range of different knowledge in an occupational domain. For example, the field tests design is referencing to a certain number of pre-specified Major competencies. To test if a test taker is competent in all Major competencies that are shaping the work, a test must contain a certain amount of items. For example, the occupational distinguishes nine major competencies. If we agree to the notion of 10-15 items per major competencies to reveal valid and reliable evidence about the competence of a test taker, a final test will contain between 90 and 120 different knowledge “units”. This extensive coverage of a knowledge domain with other item and test forms is hardly to achieve (see below).

- The construction and evaluation of multiple-choice items and tests can rely on scientifically elaborated rules and methods. For example, item construction can rely on detailed rules founded in methodical research. Moreover, is possible to analyze multiple-choice items on a very detailed level (e.g. options) by using established statistical procedures.

There are also some disadvantages associated with the use of multiple choice items. Because these disadvantages limit the range of diagnostic value of the field tests, it is necessary to quantify these factors and/or provide strategies to cope with them. Furthermore it is not the diagnostic value which counts; moreover it is the identification of unclear and false Test items:

- Test takers have to deal with symbolic representations (e.g. text, pictures; multimedia) instead of interacting with real life content (e.g. working material, and tools). Therefore, test scores depend on reading skills and the ability of test takers to reason about de-contextualized stimulus material (moderator variables of test performance). For example, if an item addresses the correct sequence of a work procedure it can happen that a test taker in practice will carry out the procedure in the correct sequence (guided by context sensitive memory structures) but fail to understand the representation of this sequence in decontextualized language based form. There are different strategies to minimize the unwanted influence of language skills on test results, some of them incorporated in standard item writing rules. For example, a basic rule is to avoid verbiage and keep items simple as possible. Another fundamental strategy to deal directly with the problem of language barriers is to develop the test in more languages (e.g. English and German). We address English language skills directly in the item field test as a moderator variable of test performance–presenting correlation coefficients that allow evaluation of the degree to which test scores depend on language skills.
- Multiple-choice items/tests are more or less blind regarding dynamic and sensomotoric aspects of work. Thus, the potential of multiple choice items to test implicit skills and experiential knowledge is extremely limited. Indeed, evidence about these aspects of work related competence only deliver working probes. Working probes, on the other side, draw vastly from resources in their development, and are subject to constraints that practically rule out the possibility to test competence in all Major competencies of an occupational standard with them. It would be ideal to use both test forms in combination, starting with the multiple choice test and ask passing test takers for a working probe. The development of working probes, therefore, is a worthwhile effort in the future for the assessment to improve the validity of testing and certification.

2.5 First Iteration of Validation / Test Item Generation

To generate an initial set of items for the Field Test, a three-day item generation workshop should be organized. The procedure and results of the workshop are further elaborated hereafter.

2.5.1 Participants

Practitioners. The most important participants in the workshop are practitioners of the occupational sector with distinguished experiential knowledge from the shop-floor level. In addition, one external subject matter expert and a number of workshop facilitators should participate in the workshop. To ensure that the items are of relevance assessing occupational competence, the participation of practitioners from the shop floor in the process of item writing is a necessity. The test developers should be subject Matter experts or higher graduates with more than 8 years of work experience, an actual position on (or close to) the shop-floor level, and sufficiently fluent in (spoken) English language.

The employers will get an official invitation letter from the TUHH to detach a person with these characteristics for a three day workshop.

In addition, the TUHH will contact company top managers via phone in order to emphasis the importance of participating practitioners for the success of the workshop. A total number of six - 10 practitioners from five - 8 companies are helpful. During the workshop, each participant will be asked to (self-) assess the degree to which he has insights into the work.

External subject matter experts. The role of this expert is to do on the spot evaluation (e.g. checking technical correctness and plausibility) of items delivered by the practitioners. In addition, he acted as a dialog partner for the practitioners – helping them to explicate their experiential knowledge about the work processes in a way that good items are harvested. The role description of the expert asks for a person with substantial technical expertise, excellent communication skills and a good general overview about the Major competencies.

Facilitators. Finally, two facilitators will be necessary to complete the workshop. One facilitator with excellent English competencies will be in charge of writing the items directly into the computer. The other facilitators will be in charge of establishing and sustaining a positive and focused item writing process in accordance with the above-stated test specifications.

2.5.2 Process of Item Generation

The item generation workshop should be conducted in 2-3 days. On the first day, the participants will introduced to each other, and to the purpose and agenda of the workshop. Particularly, the introductory part included short presentations of

- the overall concept and approach occupational Standard

- the Test Difficulty Range,
- the item format (four options multiple-choice) including item writing rules (with examples), and
- content dimensions (Major Competencies, types of cognitive performance induced by items) (with examples).

Each presentation will follow a question and answering session to ensure maximum clarity among the participants about the purpose of the workshop. Moreover, the item related presentations will use many model items to increase a common understanding of item writing rules (e.g. difference between active and receptive items). The process of item generation and writing, basically, will realized iteratively in single or group work and brainstorming and open discussion processes with all experts.

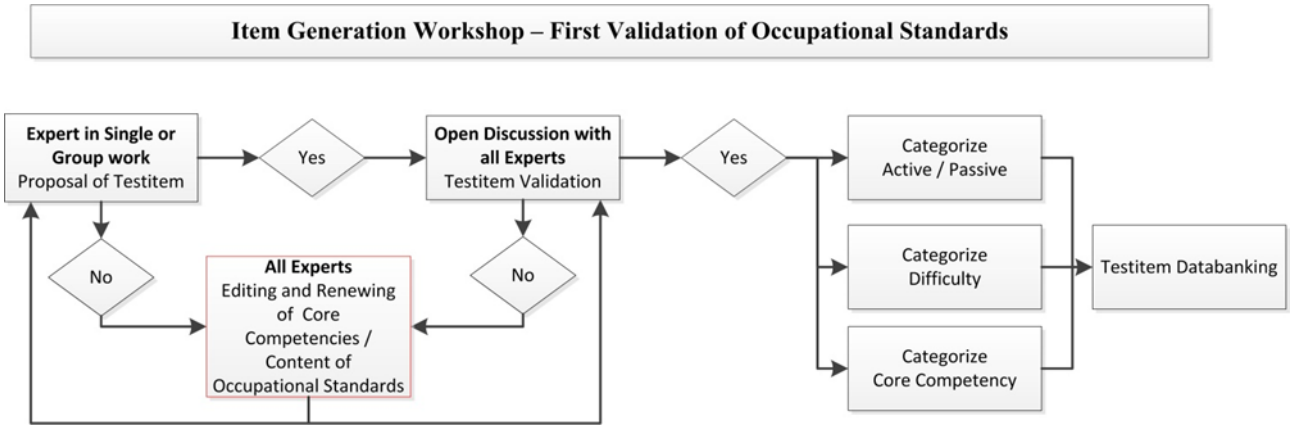
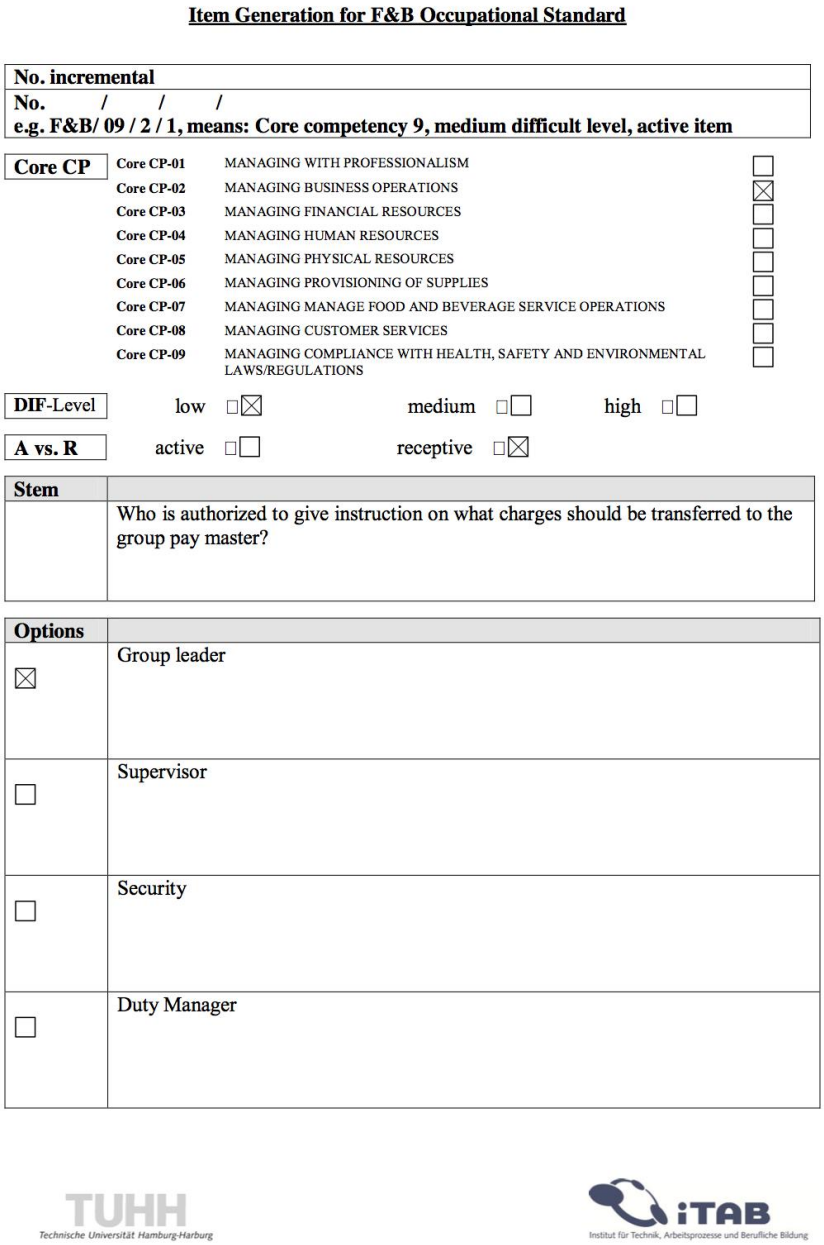


Figure 4 First Validation of an Occupational Standard

The practitioners will be encouraged to verbally express their ideas for items (stems and/or options) derived from real life situations (including a short statement why they think it is a good item). If an Item can be related to a Major competency, other workshop participants inquired or showed their agreement on the item. When a potential item started to crystallize in the discussion, it was written in the computer and projected to the group in order to support mutual fine-tuning of the emerged item. In case if it is hard or no item can be generated for Major competency the Standard needs to be improved or changed. These changes will be done by the facilitators and recorded for later purposes. Thus, if the flow of the item generation process stagnated, the moderators applied different strategies to maintain a sufficient rate of item generation. For example, a) loud reading of core work processes descriptors mentioned in the occupational standard or b) presenting pre-existing items from a German data bank and letting the practitioners decide if the specific item content is relevant for major competencies.

An obligatory step in the process of item generation is to ask the participants to judge the item difficulty for the target population, and to allocate it in an item generation matrix.



Item Generation for F&B Occupational Standard

No. incremental
No. / / / **e.g. F&B/ 09 / 2 / 1, means: Core competency 9, medium difficult level, active item**

Core CP	**Core CP-01**	MANAGING WITH PROFESSIONALISM	☐
	Core CP-02	MANAGING BUSINESS OPERATIONS	☒
	Core CP-03	MANAGING FINANCIAL RESOURCES	☐
	Core CP-04	MANAGING HUMAN RESOURCES	☐
	Core CP-05	MANAGING PHYSICAL RESOURCES	☐
	Core CP-06	MANAGING PROVISIONING OF SUPPLIES	☐
	Core CP-07	MANAGING MANAGE FOOD AND BEVERAGE SERVICE OPERATIONS	☐
	Core CP-08	MANAGING CUSTOMER SERVICES	☐
	Core CP-09	MANAGING COMPLIANCE WITH HEALTH, SAFETY AND ENVIRONMENTAL LAWS/REGULATIONS	☐

DIF-Level	low	☐☒	medium	☐☐	high ☐☐
A vs. R	active	☐☐	receptive	☐☒	

Stem	
	Who is authorized to give instruction on what charges should be transferred to the group pay master?

Options	
☒	Group leader
☐	Supervisor
☐	Security
☐	Duty Manager

TUHH Technische Universität Hamburg-Harburg

iTAB Institut für Technik, Arbeitsprozesse und Berufliche Bildung

Figure 5: Item template form to rate and judge the test items

2.6 Operation of Field Test

Generated items need to be reviewed before the field test can be conducted (cf. Schmeiser & Welch, 2006, 328). The main purpose of an item review is to minimize the risk that test takers (in particular specified groups of test takers) are not able to deliver evidence that they possess the required competence because of deficiencies or inconsistencies of test items or the test itself. The item review will use the following criteria:

- Fit within specified test dimensions (e.g. Major Competencies);
- technical and formal correctness;

- clarity (key correct, alternative responses incorrect and effective);
- Comprehensibility (from the perspective of the typical test taker).

Subject matter experts and testing experts will carry out the review. In addition, practitioners who have attended the item generation workshop will also participate in the item review. The participation of practitioners was realized by means of follow-up interviews. The practitioners will receive a confidential copy of the preliminary test (without keys) with the instruction to review the items according to the set of established criteria, thereafter, in the interview the practitioner and an TUHH member recapitulates each item and will record any feedback about debatable items.

In the final step of item review and refinement, debatable items (keys and responses) will be revised or replaced. Whenever available and appropriate, recommendations from the reviewers on how to revise or replace items will be taken into account in this process. Proofreading and final approvals will be conducted on the total set of items that have been endorsed before preparation for field-testing.

After that the Field Test delivery Through an Online Test system called ILIAS. The development of the ILIAS software started with the intention to reduce the costs of using new media in education and training. In addition, the software, according to the developers, shall ensure a maximum of efficiency in learning and testing. It is also noteworthy that the ILIAS software is running under the General Public License and is free of charge.

ILIAS offers an integrated environment for the design, delivery, and administration of tests. The test and assessment module of the system supports the development of multiple-choice items via integrated authoring tools (www.ilias.de). However, the system also supports other item formats.

Items are stored in a pool or item data bank. Thus, they are available for new test development as well for dynamic or adaptive test delivery in the future. An interesting feature (which potential for occupational testing could be explored in the future) is the possibility to use performance indicators (e.g. reaction time, wrong answers) as triggers for the provision of learning content in the test situation. All gathered data can be easily exported into conventional data analysis software (e.g. Excel or SPSS). It should be mentioned that the main purpose of ILIAS is the development and realization of web-based learning (including the management of online courses). The support of testing, therefore, is only one but important module of the system.

2.7 Second Iteration of Validation / Field Test Analysis

The analysis of how items have performed in the field test focus on classical item indices such as item difficulty and item discrimination (see exact definitions below). These classical statistics are relatively simple to compute and to understand (even for lay people). In addition, classical statistics do not re quire sample sizes as large as required by statistics derived from Item Response Theory

(IRT). However, classical statistics, in general, are less sensitive to items that discriminate differentially across different levels of ability and less likely identifying items that are statistically biased. In compliance with the recent praxis of test development, it is recommended to consult IRT statistics when more item response data is available to identify discrepancies that warrant further evaluation and refinement of test items.

It is important to notice that the interpretation of item indices depends on the purpose of eliciting test scores. In tests designed to measure whether test takers have the minimum acceptable level of knowledge and skills (criterion- referenced test) – like in the actual context field test - appropriate representing domain content and skills are more relevant than maximizing for example item discrimination (cf. Schmeiser & Welch, 2006, 338-339). For example, a test serving this purpose inevitably contains items that almost all examinees answer correctly but measure indispensable content that cannot easily be discarded from the test to be valid.

2.7.1 Item Difficulty

Item Difficulty (p-value) is the proportion of test takers answering the item correctly. The index ranges from zero to one and has an inverse meaning: the higher the p-value, the easier the item, and vice versa. Obviously, the index will vary depending on the sample responding to the item, what makes it very important that the field-test sample represents the target population.

The p-value serves as a first signal that an item does not perform well. Depending on purposes, p-values in a range of .30 to .70 are desirable to achieve a good discrimination between persons that have higher or lower abilities.

$$P_i = \frac{\sum_{v=1}^{n} x_{vi}}{n \times \max(x_i)} \times 100$$

2.7.2 Testitem Discrimination

Usually, upper and lower 25, 27, or 30 percent of scores are used to define the extreme groups. The formula of the discrimination index is Di = (Ui/NiU) – (Li/NiL). In general, a zero or negative discrimination index gives reason to consider discarding or improving an item. On the other side, a discrimination index of .25 or larger is a signal that the item does not need improvement Upper-lower 25% scores D-values in the sample have a mean of .30 and a standard deviation of .2.

$$D = x_v^+ - x_v^-$$

2.7.3 Testitem Correlation

In addition, correlation between item scores and total test scores (biserial or point-biserial correlations for items scored dichotomously) are also used to evaluate the items.

$$r_{it} = (x_{vi}, x_v)$$

Depending on the field sample and the test purpose, it will be decided to accept items with a zero or slight negative correlation. Taking into consideration traditional indices the overall item performance is acceptable for the test purpose.

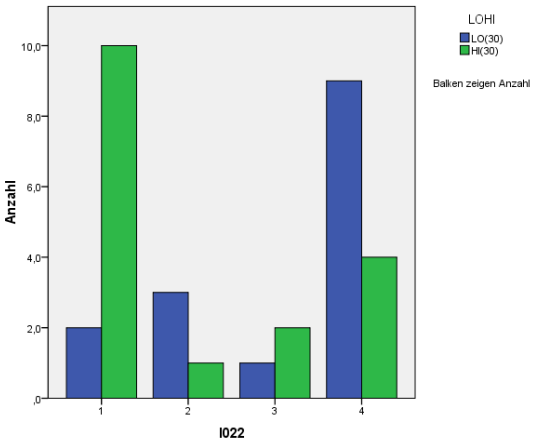
2.7.4 Internal Consistency of Testitems

Cronbach's coefficient alpha will be used to compute the internal consistency reliability estimate. This coefficient eliminates a potential error source associated with reliability estimates based on divisions of the test in two or more parts that are insufficiently parallel in content and difficulty (Haerte, 2006, 73). Cronbach's alpha increases as a function of item interrelatedness and test length. Cronbach's alpha is not a measure of one-dimensionality of a set of items although this is often assumed in the literature. Internal consistency is necessary for homogeneity, but it is not sufficient (cf. Schmitt, 1996). The index gives only a hint, whether the Test item fits to the Major competencies or not, of better, whether the test items have the same dimension to proof.

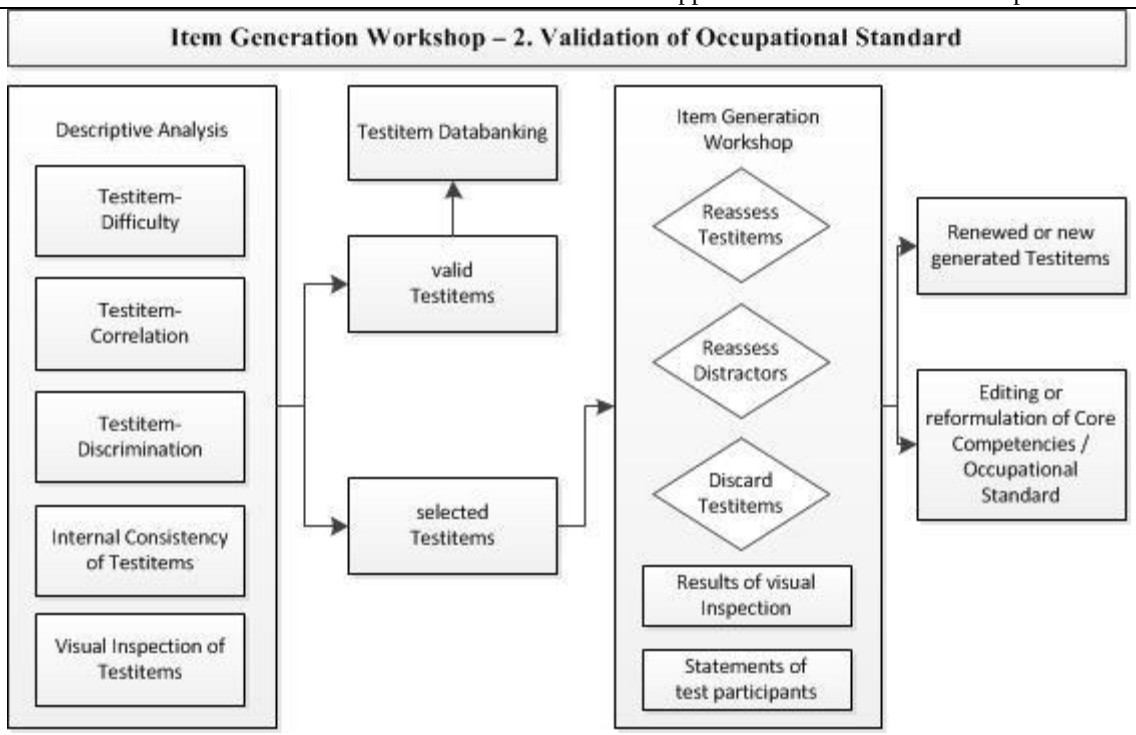
$$\alpha_{st} = \frac{N \times \bar{r}}{1 + (N - 1) \times \bar{r}}$$

2.7.5 Visual Inspection of Testitems

The visual Inspection of Items is also very helpful to identify weaknesses in test items and to see the range of key answers and distractors of Hi-Performers and Low Performers.



2.7.6 Second Item Generation Workshop



The second validation is important to judge and rate the quality of the selected test items. The results of the test item analysis can give hints of the quality of the occupational standard. During the first validation the expert rated the test item to be valid. After the field test and the analysis, a lot of test items needs to be reassessed.

During a second, much shorter, item generation workshop the test items will be discarded, reassessed or the distractor needs to be refined. The discussion process with the expert in the item generation workshop about the results of the field test will help to improve the quality of both, the test items and the standards.

3 Summary

The validation process has positive effects to occupational standards:

- The reliability of test items due to identification of item difficulty and discrimination
- The approved relationship of the test items to the Major competencies of the occupational standard
- The representative status of an occupational standard to its work environment
- Validated test items can be used for future competence tests; the test items have proofed their correct test dimension in the field test.
- In two different iterations the occupational standard will be validated

Anastasi, A. (1988). Psychological Testing. New York, New York: MacMillan Publishing Company

Anderson, L.W. & Krathwohl, D. (Eds.). (2001). A taxonomy of learning, teaching, and assessing: A revision of Bloom/s taxonomy of educational objectives. New York: Longman.

Bond, L.A. (1996). Norm- and criterion referenced testing. Practical Assessment, Research, & Evaluation, 5(2), Retrieved June 29, 2009 from http://PAREonline.net/getvn.asp?v=5&n=2.

Bloom, B.S. (Ed.). (1956).Taxonomy of educational objectives: Book 1, Cognitive domain. New York: Longman.

EUROPEAN CENTRE FOR THE DEVELOPMENT OF VOCATIONAL TRAINING (CEDEFOP): The dynamics of qualifications. Defining and renewing occupational and educational standards. (Cedefop panorama series, Bd. 176). Luxembourg 2009.

Clauser, B.E. & Case, S.M. (2006). Testing for Licensure and Certification in the Professions. In R.L. Brennan (Ed.), Educational Measurement (4th ed., pp. 701-731). Westport: American Council on Education, Praeger Publishers.

GIELEN, P.M/REITSMA, Nicky/CUNNINGHAM-BROWN, Wendy: Towards a competent labour force. Development of and experiences with competence-based education. Wageningen 2000.

Haladyna, T.M. & Downing, S.M. (1989). A taxonomy of multiple-choice item writing rules. Applied Measurement in Education, 2, 37-50.

Haladyna, T.M. (2004). Developing and validating multiple choice test items (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.), Educational Measurement (4th ed., pp. 65-110). Westport: American Council on Education, Praeger Publishers.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), Technology in mental health care delivery systems (pp. 119-l37). Norwood, NJ: Ablex.

Schmeiser, C.B. & Welch, C.J. (2006). Test Development. In R.L. Brennan (Ed.), Educational Measurement (4th ed., pp. 307-353). Westport: American Council on Education, Praeger Publishers.

Schmitt, N. (1996). Uses and Abuses of Coeffcient Alpha. Psychological Assessment, 4, 350-353.

Zsambok, C.E. and Klein, G (1997) Naturalistic Decision Making. Lawrence Erlbaum Associates, Mahwah, NJ.

Lave, J. & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. New York: Cambridge University Press.