How do I develop a psychological test or questionnaire?

Abstract

The purpose of this *How-to article* is to provide physicians and other health professionals working in the field of medical education research with a basic understanding of the construction of tests or questionnaire measures. The construction of such measures is too complex to be described on a few pages. Therefore, this article can only enable readers to roughly evaluate such measures or to convey an idea of how these are generally constructed.

The article outlines various phases of test or questionnaire construction. It begins with the *content phase*, in which a construct is defined, if possible, by drawing on theories and models. Here, items are written, a response format is selected, the instruction is formulated, and pilot tests are conducted. In the *structural phase*, the structure of the test or questionnaire is evaluated using suitable test statistical methods and statistical parameters. In the final phase (*external phase*), additional evidence for the validity of test or questionnaire results is sought. The validation of such measures is not the last step in the construction of tests or questionnaires as it is to be considered in all phases of test or questionnaire construction. The validation of test and questionnaire measures is theoretically and methodically demanding and should never be considered complete. Strictly speaking, it should not be said that a test or questionnaire is valid, because validity is not a property of such measures. It rather is statements and conclusions based on test or questionnaire results that can be valid.

Keywords: phases of test- and questionnaire construction, reliability, validity, generation and wording of items

Marianne Giesler[1]
Götz Fabry[2]

1 Freiburg i. Brsg., Germany

2 University Freiburg, Department of Medical Psychology and Medical Sociology, Freiburg i. Brsg., Germany

1. Goal of this how-to article

In *medical education* research, tests and questionnaires are often used, for example, to measure motivation, empathy, or certain performance levels of students. However, the training of physicians and medical professionals hardly conveys the competencies that would enable them to evaluate the quality of such measures, develop measurement instruments, or translate an existing questionnaire. This *how-to article* is intended to explain and illustrate the development of psychological test and questionnaire measures.

The process of test or questionnaire construction is complex and time-consuming. There are various specialised books in which this process is described in detail, usually on more than 200 pages [1], [2]. A short article can therefore only enable readers to roughly evaluate tests and questionnaires and convey an idea of how such measures are generally constructed according to the so-called *Classical Test Theory* (CTT) (see below).

The following paragraphs will first briefly explain the different types of tests and questionnaires. Then, the various phases of constructing such measures will be outlined.

2. What types of test or questionnaire measures can be distinguished?

Psychological test and *questionnaire measures* can be assigned to three areas: performance tests, personality questionnaires, and projective techniques [3]. Each area can be further subdivided (see table 1). These measures may cover abilities, skills, characteristics, and states of persons that often are not directly observable, but derived from observable behaviours, and referred to as constructs. Well-known psychological constructs used in *medical education research* are, for example, motivation, self-efficacy, resilience, reflective ability, and empathy. Since constructs cannot be directly measured, they are referred to as latent variables, for which items are used as indicators [2].

Most psychological test and questionnaire measures are based on the assumptions and construction principles of the so-called *Classical Test Theory* (CTT) [4], [5], which assumes that individual measurements can vary across different points of measurement. Its basic concept involves the assumption that the observed value X of a

Table 1: Psychological test and questionnaire measures – an overview with examples [3]

Performance tests (abilities/skills of a person)	**Personality tests** (traits/attitudes of a person)	
	Questionnaires	**Projective techniques**
Developmental tests (e.g. Eggenberger counting test) Intelligence tests (e.g. Intelligence Structure Test) General performance tests (e.g. d2-Test) School tests (e.g. school readiness tests, math tests) Special functional assessment and aptitude tests (e.g. Inventory of Social Competencies)	Job application questionnaires (e.g. test for medical study programs) Interest tests (e.g. career interest tests) Clinical tests (e.g. Beck Depression Inventory) etc.	Form interpretation methods (e.g. Rorschach Test) Verbal-narrative methods (e.g. Thematic Apperception Test) Drawing methods (e.g. Tree Test) Designing methods (e.g. Thematic-Design-Test)

person on a test consists of both the person's *true score* and a *random measurement error*. The result of an intelligence test would accordingly be influenced by the actual intelligence of the person being tested and by unsystematic influences, such as performance fluctuations due to the time of day, e.g. if one were to conduct an infinite number of measurements, the mean of these measurements would correspond to the person's actual intelligence score.

In addition to the CTT, there is the *Probabilistic Test Theory* (PTT), which is sometimes also called *Item-Response Theory* (IRT). This theory assumes that the probability of a specific response to an item depends on the characteristics of the item and the level of the latent trait being measured in the person [5]. According to Rost [4], the two test theories CTT and PTT are not, as often described, competing, but complementary methods, since one theory starts where the other ends, or because both test theories are largely based on the same assumptions. Further details on PTT can be found in Bühner [2] and Döring and Bortz [5].

3. How are test and questionnaire measures developed?

When developing a test or a questionnaire measure, so-called test quality criteria must be fulfilled (see table 2). The development of such measures begins with the determination or definition of the construct to be measured. After that, items (tasks or statements) are constructed, and the answer format is selected. After a pretest, the measure is specifically tested on one or more samples. If a sufficiently large number of data has been obtained, it is analysed how reliable the test or questionnaire measures the construct (reliability) and whether it measures the construct it claims to measure (validity).

3.1. Definition of the construct

To define and operationalize the construct, theories or models are used, if available. Examples of constructs based on sound theories and models that have been used to develop psychological tests include motivation and learning strategies. If theories and models are not available, the construct space can be narrowed down after extensive literature study, and indicators of the construct (e.g. specific statements or behaviours) can be determined. A current example from medical education research where such a procedure is necessary is *reflective ability*. There are various models and theories here as well, but they differ significantly in what is understood by *reflective ability*. Therefore, to develop a test procedure for *reflective ability*, it would first be necessary to define which indicators of *reflective ability* should be considered based on prior work. As part of the construct definition, it should also be determined to what extent relationships and overlaps with other constructs exist (nomological network) [2]. For example, there has been an illustrative discussion as to the extent to which the personality trait of *openness to experience* is related to creativity [6].

The quality of the definition of the construct determines how easily items can be generated. A detailed definition considering necessary distinctions from other constructs also increases the likelihood of the *content validity* of the construct [1], [2].

3.2. Generation and wording of items

Different sources can be used to generate items [1]. For example, items can be

- derived from *theories* or from an extensive, systematic review of the *literature,*
- generated from the results of *preliminary investigations* (interviews, focus group discussions, etc.),
- written in accordance with *existing tests and questionnaires,*
- developed by *experts*.

Table 2: Traditional quality criteria of tests [2], [5], [10], [19]

Quality criteria[1]	Definition and types of quality criteria[1]	
Objectivity	Objectivity refers to the degree to which the results of a test are independent of the conditions of test administration, scoring, and interpretation.	**Objectivity in the administration of a test**: The administration of a test should always be the same for each tested person or comparable. **Objectivity in the scoring of a test**: The determined test score is independent of the person scoring the test. **Objectivity in the interpretation of test results**: The interpretation of the determined test score is independent of the person performing the interpretation.
Reliability	The reliability of a test indicates the degree of measurement accuracy of an obtained score, i.e. how strongly this score is influenced by random errors of measurement. To determine reliability, the methods described in the adjacent column are possible, which are based on the principle of multiple measurements of a characteristic.	**Test-retest reliability** (stability of the measured characteristic): Correlation of test scores obtained at two different occasions. **Parallel-forms reliability** (stability of conditions): Correlation of test scores obtained through two different tests measuring the same characteristic. **Split-half-reliability**: Correlation of the scores obtained in one test half with the scores obtained in the other one. **Internal consistency** (measurement accuracy at a specific measurement time): Correlation of each individual test item with every other item measured at the same time, corrected for test-length. One of the most common indicators of internal consistency is Cronbach's α.
Validity	Validity indicates whether the test measures what it claims to measure. Its main point is to convincingly demonstrate that the conclusions drawn from a person's test score in regard to a behaviour or characteristic are appropriate.	**Content validity** (judged by experts): The items of a test cover a representative sample of the manifestations of the attribute to be measured. **Face validity** (judged by non-experts): Lay persons and test takers perceive the items of a test as a representative sample of the characteristics of the attribute to be measured. **Construct validity**: The test correlates meaningfully with other constructs as predicted by hypotheses based on test-content or theoretical considerations. **Criterion-related validity**: The test correlates positively with corresponding content-related characteristics measured outside the testing situation.
Selected secondary quality criteria	**Comparability**: If a test is to be taken by a person multiple times or if one intends to bar respondents from copying responses in group testing situations, it can be advantageous to present a test in parallel forms or tests of comparable validity. **Economy** of a test is given when it requires relatively few resources, such as time, money, etc., in relation to the diagnostic insight gained. **Usefulness**: A test is useful if it measures or predicts a characteristic for which there is a practical need. **Fairness** A test is said to be fair when, for example, people of equal ability have the same chance of receiving a test score independent of age, gender, or culture.	

[1] Publications by Loevinger [9] and Messick [13] suggest an alternative representation of the validity criterion, with a more suitable conceptual model. However, in this article, we agree with Döring and Bortz [5], who have chosen a *strategy of deliberately imprecise wording* to better convey understanding test quality criteria.

When generating items, the goals of the test being constructed should be considered [2]. If the goal is to capture the trait or ability manifestations of individuals, content-valid items should be constructed. A test for detecting *fear of progression*, i.e. the fear a diagnosed condition might progress and deteriorate, is valid in terms of content if the test items can be considered a representative sample of the entire range of *fear of progression* (e.g. cognitive, emotional, and behavioural aspects). It should be ensured that only one construct is captured with the items. Furthermore, all indicators of a construct should correlate with each other [2].

To ensure the *content validity* of the test, attention should be paid to collect a sufficiently large and representative number of items. The number of items in the drafted test should be greater than the planned number of items in the final version [2].

Before constructing the items, it should be decided how exactly items should be written. For example, this can be done in the following ways:

- As *questions*: Do you feel respected by members of other health professions?
- As *statements*: I feel respected by members of other health professions.
- In the *first person singular*: I enjoy working with members of other health professions.

- In an *impersonal form*: People enjoy working with members of other health professions here.

The items should be coherent and understandable in terms of content [1], [2]. This means, among other things, that foreign words or complex sentence structures should be avoided. The items should also be clearly defined in terms of content. For this purpose, if possible, avoid conditional statements or conjunctions, among other things. Negations (especially double negatives) should also be avoided.

3.3. Choosing the response format

The selection of appropriate response options is just as important as constructing the items. Frequently, psychological test and questionnaire measures use rating scales (usually so-called Likert scales), with graded response categories to which verbal labels are attached. Labels often encountered range from "not applicable" to "applicable" or "very poor" to "very good". Rating scales may also differ in the number of response categories. In this regard, response scales with up to 7 levels are acceptable [2]. Furthermore, it must be decided whether the response levels of the items are unipolar (e.g. "never" to "very often") or bipolar (e.g. "disagree", "slightly disagree", "neither disagree/nor agree", "slightly agree", "agree"). In addition to verbal labels of the response levels, visual aids can also be used (e.g. smileys).

3.4. Wording of the instruction

The purpose of the instruction is to familiarize respondents with the content and purpose of the test or questionnaire measure, provide guidance on how to answer the items, and explain data protection regulations [7]. It has a central function, as it not only prepares for the task of taking the test, but can also create a pre-set attitude in the people being questioned about the task to be completed [1]. An instruction is usually drafted at the end of the construction process, after the items and response alternatives have been determined. In addition to specifying the objective or purpose of the test or questionnaire, instructions usually contain information indicating that

- participation is voluntary and that there are no disadvantages to be feared in case of non-participation,
- all items should be read and answered quickly,
- the items are to be responded to one after the other and no item should be skipped, even if this may seem difficult at times, and that in this case the "most likely" option should always be checked,
- confidentiality and anonymity of individual information is ensured in accordance with applicable data protection regulations.

3.5. Conducting preliminary tests

Conducting one or more pretests is another important prerequisite for the development of a test or questionnaire measure. However, there are no generally accepted procedural rules for carrying these out. For example, recommendations vary greatly when it comes to determining the number of cases necessary for this [8]. However, a small number of individuals are usually asked to provide feedback on the comprehensibility of the items and instructions, and to report any difficulties encountered while completing the measure. It is important that these individuals are as similar as possible to the subsequent target group of the test or questionnaire, e.g. in terms of language comprehension. Preliminary tests also provide information about the time needed for completion, the respondents' interest in the topic, and the possible distributions of the responses. Based on the feedback, the measure will be modified if necessary.

4. Statistical evaluation of psychological test and questionnaire measures

The process of statistically evaluating a test or questionnaire measure can be subdivided in accordance with phases outlined by Loevinger [9], as follows:

- *Substantive phase*: During this phase, the measure is theoretically grounded and based on available literature. Pretests are conducted to clarify the comprehensibility of the items and problems with answering them.
- *Structural phase*: The primary focus of this second phase is on examining the structural (e.g. factorial structure) and further psychometric properties (e.g. item correlations) of the measure.
- *External phase*: In this phase, the extent of the agreement of the measure with other criteria and, if applicable, similar tests or questionnaires should be determined.

All previous descriptions in this *how-to article* can be assigned to the *substantive phase* (see table 3). The following sections focus on the psychometric analysis of test or questionnaire measures that are assigned to the other two phases.

4.1. Structural phase

In the *substantive phase*, the *face* and *content validity* of test or questionnaire measures can already be ensured. However, the structural and psychometric properties of test or questionnaire measures can only be determined after the test and questionnaire measure has been taken by individuals from the respective target group (data collection). First, a dimensional analysis should be performed using factor analyses (statistical methods that group the variables according to their intercorrelation; *factorial validity*), followed by determining the test's reliability and an item analysis [10]. However, if the sample size is too small [2] for dimensional analyses, preliminary reliability calculations can be conducted and the items can be analysed regarding their difficulty, discriminant

Table 3: Phases of constructing test and questionnaire measures (see chapter 4)

Phases	Steps	Relevant aspects	Test quality criteria
Substantive phase	Definition of the construct	– theories, models – literature review – nomological networks	
	Generating and wording of items	Generating items theories, literature, preliminary research (e.g. focus groups), existing tests, experts Item wording question form, statements, first person singular, impersonal form, no complicated sentence form, no foreign words, no double negation, etc.	face validity content validity
	Deciding on a response format	– rating scales (usually Likert-type) – design of rating scales: unipolar, bipolar, visual aids (e.g. smileys)	
	Writing of instruction	– purpose of the test/survey – state that all items should be read and answered quickly – state that items should be answered one by one, without leaving out any item – give assurance of anonymity and add notes on data protection – etc.	
	Conducting pretests	– to obtain feedback on the comprehensibility of the items and instructions – to obtain information about difficulties in answering items – in order to capture the required implementation time – etc.	
Structural phase	Review of the test structure	– clarifying the dimensionality of the test using confirmatory or exploratory factor analyses – determining reliability – Item analyses (M, SD, difficulty index, discriminatory power)	structural validity reliability
External phase	Gathering evidence of the validity of test scores	– establishing construct and criterion-related validity – testing hypotheses and conclusions based on the test scores	Types of validity, e.g.: - convergent - discriminant - retrospective - concurrent - group differences

validity, and intercorrelations (item analyses) (see table 4).

Recommendations for the sample size required for factor analyses vary greatly in the relevant literature. According to MacCallum et al. [11], common rules of thumb are problematic because the required sample size depends on the number of items per factor and the degree of communality (the proportion of variance of a variable that is explained by the factors) of each item. However, communalities are usually not known in advance. Therefore, in spite of the aforementioned issues, it may be mentioned here for rough orientation that it has been recommended to include a number of respondents in factor analyses that is at least five to ten times the number of items.

If the sample size is sufficient for conducting factor analyses and a hypothesis or model for the dimensions of the test is available, a *confirmatory factor analysis* should be conducted. If there are no reasonable assumptions

Table 4: Description of test statistics

Test statistics	Description
Mean	The arithmetic mean is the sum of a set of values divided by the number of values in that set.
Standard deviation	The standard deviation is a measure of the variation of the values of a variable around its mean.
Difficulty index	The difficulty index is a measure of how difficult or easy an item or a task is for a sample of respondents to respond to or solve.
Discriminatory power	The concept of discriminatory power of an item describes how well that item differentiates between individuals with low and high test scores.
Communality of an item	The proportion of the total variance of a variable or a single item that can be explained by the set of all factors specified by a factor analysis.

about the relationships between the items, an *exploratory factor analysis* is recommended.

4.2. External phase

The validation of test and questionnaire measures is theoretically and methodically demanding and should never be considered complete [5], [12]. Therefore, strictly speaking, it should not be said that a test or questionnaire is valid, since validity is not a property of tests or questionnaires (see 4.2.2). Only statements and conclusions based on test or questionnaire scores can be more or less valid.

The validation of test and questionnaire measures (or more precisely, of test or questionnaire scores) involves a variety of aspects. In this regard, however, the understanding of which indicators can be considered as signs of validity has changed over time. The traditional concept of validity is presented first, followed by the validity approach of Messick [13], which complements the traditional approach.

4.2.1. Construct and criterion validity

First, it can be determined whether the construct captured by the test or questionnaire measure correlates with other theoretical constructs in terms of content and theory (*construct validity*) and/or whether the test or questionnaire scores correlate positively with behavioural manifestations outside of the testing situation (*criterion validity*) [5].

To determine *construct validity,* additional measurement instruments can be used that capture either construct-related or construct-unrelated characteristics. According to Campbell and Fiske [14], in the first case *convergent validity* would be checked and *discriminant validity* in the second. *Construct validity* also includes the previously described *factorial validity* (see 4.1). Furthermore, it is possible to analyse differences in the test results of selected groups. That is, differences in test scores of various groups (e.g. differing by age, socioeconomic status, or education) are postulated based on theoretical considerations and empirical findings [10]. If these differences are found as predicted, they will be interpreted as evidence of validity.

In terms of *criterion validity*, several types of validity can be distinguished depending on the time of measurement of the external criterion [5]. *Retrospective validity* is checked when a criterion (e.g. past school grades) has been collected before the test scores to be validated (e.g. school performance test) is applied. In *concurrent validity*, the criteria (e.g. complaints in medical consultations such as sleeplessness and listlessness) are recorded (almost) at the same measurement time as the test scores to be validated (e.g. results of a measuring instrument for recording the extent of depression). In *predictive validity* the criterion score (e.g. academic performance) is recorded later than the test score to be validated (e.g. results of a medical college admission test). Determining *criterion validity* requires that the chosen external criterion is reliable and valid.

Incremental validity is also a type of *criterion validity*, but it is rarely tested. If *incremental validity* is analysed, an established test or questionnaire measure is used that claims to measure the same characteristic as the measure to be validated. The new measure should then significantly improve the prediction of the external criterion [5].

4.2.2. Argument-based validation concepts

The classical concept of validity described in the previous section was expanded by Messick [13]. He describes six general validity aspects, which apply to all diagnostic measurements in the educational sector. They are based on the fundamental idea that the validity of a diagnostic measurement cannot be considered solely as a numerical coefficient, but rather as a theoretically and empirically founded argument for the validity of test score interpretations. In other words, “it is incorrect to use the unqualified phrase *the validity of the test*” ([15], p.11), because the observed test scores are not only a function of the items but also depend on the respondents and the context of the evaluation [13]. Validity can therefore be understood as an argument for the validity of the interpretation of test scores based on evidence regarding these six aspects. In table 5, the validity aspects described by Messick are presented. It becomes clear that only the aspects of *substantive validity, generalizability* and of *consequential validity* supplement the traditional approach (see table 5).

Table 5: Validity approach by Messick [5], [13], [20]

Aspects of validity	Meaning
Content	Evidence of content relevance, representativeness, and technical quality of the construct.
Substantive	Not only should the content of a diagnostic procedure be representative of the feature domain to be captured, but also the cognitive processes needed to answer the test questions. That is, the test subjects should primarily focus on the content of the test (and not on understanding a complicated formulation of the test instructions or an item). The correspondence between the cognitive processes that occur in the testing situation and the ones assumed to be central for the construct being tested should be empirically testable. This can be done, for example, during piloting the test in the pretest phase, using the "method of thinking aloud" or by analysing eye movements.
Structural	Convergence of the theoretical model or the construct structures with the empirical measurement model: The structure of test data determined empirically by means of factor analysis, for example, should be consistent with the assumptions about the structure of the construct (e.g. aspects of motivation, intelligence, etc.).
Generalizability	Consideration of the extent to which the properties and interpretations of the test scores can be generalized to other tasks, population groups, and settings.
External	Refers to relationships between test scores and other characteristics (convergent, discriminant validity, etc.) that are found to be consistent with prior hypotheses.
Consequential	Consideration of both the positive and negative consequences of using test scores. *Teaching-to-the-Test* would be an example of an undesirable consequence of using a school performance test. In this case, content and methods of teaching would focus on preparing for the upcoming test.

Additionally, Messick [13] pointed out two potentially confounding variables that could affect validity. A construct may be *underrepresented* because it is too narrow and does not cover important dimensions or facets of the construct. This would be the case, for example, if a test of performance anxiety only captures its emotional component and disregards its cognitive and physiological components. However, validity can also be limited by *construct-irrelevant variance*, if test items are too difficult or too easy for some individuals [13]. This is the case, for example, when the correct completion of tasks in a mathematics test also depends on its unreasonably high demands on the respondents' language comprehension. These expansions of the classical concept of validity have by now been adopted by, among others, the *American Educational Research Association* (AERA) and the *American Psychological Association* (APA) [15], [16].

5. Translation of a test or questionnaire measure

In the past, tests were often translated using the forward-backward-translation method. That is, the test was first translated into the target language, then this translation was re-translated [17] by another person, and then the original and the backward-translated versions were compared. However, a simple backward translation cannot eliminate all translation problems, so multi-stage translation processes are now recommended [17]. For example, according to the *European Social Survey Programme* for translating questionnaires, a five-step translation framework called *TRAPD* is suggested. This acronym stands for *T*ranslation, *R*eview, *A*djudication (deciding on a version), *P*re-testing, and *D*ocumentation [18]. These five steps should be taken in a team effort from the beginning. A complete statistical evaluation of the translated version is also required when translating a test.

6. Summary

The construction of test or questionnaire measures requires a well-defined construct or at least a clearly described construct space. Based on this, items can be written that must be content-valid and easy to understand and that are oriented toward the goals of the measure. If the measure has been supported in pretests with small groups of people, its structural (dimensionality) and further psychometric (reliability, validity, etc.) properties can be checked using more extensive data collections. To determine the validity of the test results, various aspects need to be considered. These relate primarily to the construct to be measured and its theoretical embedding as well as to its relationship to other variables, but also to the context of the measurement and the consequences derived from the test results.

Authors' ORCIDs

- Marianne Giesler: [0000-0001-9384-2343]
- Götz Fabry: [0000-0002-5393-606X]

Competing interests

The authors declare that they have no competing interests.

References

1. Mummendey HD, Grau I. Die Fragebogen-Methode. 5. Aufl. Göttingen: Hogrefe Verlag; 2008.
2. Bühner M. Einführung in die Test- und Fragebogenkonstruktion. 3. aktual. u. erw. Aufl. München: Pearson Studium; 2011.
3. Brähler E, Holling H, Leutner D, Petermann F. Brickencamp Handbuch psychologischer und pädagogischer Tests. 3. Aufl. Göttingen: Hogrefe; 2002.
4. Rost J. Was ist aus dem Rasch-Modell geworden? Psych Rundsch. 1999;50(3):140-156. DOI: 10.1026//0033-3042.50.3.140
5. Döring N, Bortz J. Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5. vollst. überarb., aktual. u. erw. Aufl. Berlin: Springer; 2016. DOI: 10.1007/978-3-642-41089-5
6. King LA, Walker LM, Broyles SJ. Creativity and the five-factor model. J Res Pers. 1996;30(2):189-203.
7. Reinders H. Fragebogen. In: Reinders H, Ditton H, Gräsel C, Gniewosz B, editors. Empirische Bildungsforschung. Strukturen und Methoden. Wiesbaden: VS Verlag für Sozialwissenschaften; 2011. p.53-65. DOI: 10.1007/978-3-531-93015-2_4
8. Porst R. Im Vorfeld der Befragung: Planung, Fragebogenentwicklung, Pretesting. ZUMA-Arbeitsbericht, 1998/02. Mannheim: Zentrum für Methoden und Analysen (ZUMA); 1998. URN: urn:nbn:de:0168-ssoar-200484
9. Loevinger J. Objective tests as instruments of psychological theory. Psychol Rep. 1957;3(3):635-694. DOI: 10.2466/pr0.1957.3.3.635
10. Lienert GA. Testaufbau und Testanalyse. 2. durchges. u. verb. Aufl. Weinheim: Beltz; 1961.
11. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample Size in Factor Analysis. Psychol Method. 1999;4(1):84-99. DOI: 10.1037/1082-989X.4.1.84
12. Repke L, Birkenmaier L, Lechner CM. Validity in Survey Research - From Research Design to Measurement Instruments. Mannheim: GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines); 2024. DOI: 10.15465/gesis-sg_en_048
13. Messick S. Validity of Psychological Assessment. Validation of Inferences from Persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741-749. DOI: 10.1002/j.2333-8504.1994.tb01618.x
14. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56(2):81-105.
15. AERA, APA, NCME. Standards for Educational and Psychological Testing. Washington: American Psychological Association (APA); 2014.
16. Schaper N. Validitätsaspekte von Kompetenzmodellen und -tests für hochschulische Kompetenzdomänen. In: Musekamp F, Spöttl G, editors. Kompetenz im Studium und in der Arbeitswelt. Nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen. Frankfurt, M: Lang; 2014. p.21-48.
17. Su CT, Parham LD. Generating a valid questionnaire translation for cross-cultural use. Am J Occup Ther. 2002;56(5):581-585. DOI: 10.5014/ajot.56.5.581
18. European Social Survey. ESS Round 11 Translation Guidelines. London: ESS ERIC Headquarters; 2022. Zugänglich unter/available from: https://www.europeansocialsurvey.org/sites/default/files/2024-08/ESS_R11_Translation_Guidelines.pdf
19. Moosbrugger H, Kelava A. Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger H, Kelava A, editors. Testtheorie und Fragebogenkonstruktion. 3. vollst. neu bearb., erw. u. akt. Aufl. Heidelberg: Springer; 2020. p.7-26. DOI: 10.1007/978-3-540-71635-8_2
20. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-837. DOI: 10.1046/j.1365-2923.2003.01594.x

Corresponding author:
Marianne Giesler
Freiburg i. Brsg., Germany
Dr_M_Giesler@t-online.de

Please cite as
Giesler M, Fabry G. How do I develop a psychological test or questionnaire? GMS J Med Educ. 2026;43(1):Doc9. DOI: 10.3205/zma001803, URN: urn:nbn:de:0183-zma0018034

This article is freely available from
https://doi.org/10.3205/zma001803

Received: *2025-01-20*
Revised: *2025-05-13*
Accepted: *2025-07-28*
Published: *2026-01-15*

Wie entwickle ich ein psychologisches Test- oder Fragebogenverfahren?

Zusammenfassung

Ziel dieses *Gewusst-wie-Artikels* ist es, Ärztinnen und Ärzte sowie anderen Gesundheitsfachkräften, die im Bereich Medizinische Ausbildungsforschung tätig sind, ein Grundverständnis der Konstruktion von Test- oder Fragebogenverfahren zu vermitteln. Die Konstruktion solcher Verfahren ist insgesamt zu komplex, um sie auf einigen wenigen Seiten zu beschreiben. Daher kann dieser Beitrag Leserinnen und Leser lediglich in die Lage versetzen, solche Verfahren grob zu bewerten, bzw. eine Vorstellung davon zu vermitteln, wie solche Verfahren im Allgemeinen konstruiert werden.

Der Beitrag skizziert verschiedene Phasen der Test- bzw. Fragebogenkonstruktion. Er beginnt mit der *inhaltlichen Phase*, in der ein Konstrukt nach Möglichkeit mit Rückgriff auf Theorien und Modelle definiert wird. Hier werden Items formuliert, ein Antwortformat ausgewählt, die Instruktion formuliert und Vortests durchgeführt. In der *strukturellen Phase* wird die Struktur des Tests bzw. Fragebogens mittels geeigneter teststatistischer Verfahren und Kennwerte überprüft. In einer letzten Phase (*externe Phase*) werden weitere Belege für die Validität von Test- bzw. Fragebogenergebnissen gesammelt. Die Validierung solcher Verfahren stellt jedoch keinen abschließenden Schritt der Test- bzw. Fragebogenkonstruktion dar, denn sie wird in allen Phasen der Test- bzw. Fragebogenkonstruktion berücksichtigt. Die Validierung von Test- und Fragebogenverfahren ist theoretisch und methodisch anspruchsvoll und sollte nie als abgeschlossen betrachtet werden. Es sollte strenggenommen auch nicht davon gesprochen werden, dass ein Test bzw. Fragebogen valide ist, da Validität keine Eigenschaft solcher Verfahren ist. Es können nur Aussagen und Schlussfolgerungen valide sein, die auf der Grundlage von Test- bzw. Fragebogenergebnissen getroffen werden.

Schlüsselwörter: Phasen der Test- bzw. Fragebogenkonstruktion, Reliabilität, Validität, Generierung und Formulierung von Items

Marianne Giesler[1]
Götz Fabry[2]

1 Freiburg i. Brsg., Deutschland

2 Universität Freiburg, Institut für Medizinische Psychologie und Medizinische Soziologie, Freiburg i. Brsg., Deutschland

1. Ziel dieses Gewusst-wie-Artikels

In der *medizinischen Ausbildungsforschung* werden häufig Test- oder Fragebogenverfahren eingesetzt, z.B. um Motivation, Empathie oder auch bestimmte Leistungen von Studierenden zu messen. In der Ausbildung von Ärztinnen, Ärzten und medizinischen Fachkräften werden allerdings kaum Kompetenzen vermittelt, die es erlauben, die Qualität solcher Verfahren zu bewerten, Messinstrumente zu entwickeln oder einen bestehenden Fragebogen zu übersetzen. Dieser *Gewusst-wie-Artikel* soll daher die Entwicklung psychologischer Test- und Fragebogenverfahren erläutern und anschaulich machen.

Der Prozess der Test- bzw. Fragebogenkonstruktion ist komplex und zeitaufwendig. Es gibt verschiedene Fachbücher, in denen dieser Prozess ausführlich auf meist über 200 Seiten beschrieben wird [1], [2]. Ein kurzer Artikel kann daher Leserinnen und Leser lediglich in die Lage versetzen, Test- und Fragebogenverfahren grob zu bewerten, und ebenso eine Vorstellung vermitteln, wie solche Verfahren nach der sog. *klassischen Testtheorie* (s.u.) im Allgemeinen konstruiert werden.

In den folgenden Kapiteln wird zunächst kurz erläutert, welche Arten von Test- bzw. Fragebogenverfahren sich unterscheiden lassen. Anschließend werden verschiedene Phasen der Konstruktion solcher Verfahren skizziert.

2. Welche Arten von Test- bzw. Fragebogenverfahren lassen sich unterscheiden?

Psychologische Test- bzw. Fragebogenverfahren lassen sich drei Bereichen zuordnen: Leistungstests, Persönlichkeitsfragebögen und Projektive Verfahren [3]. Jeder Bereich kann wiederum weiter unterteilt werden (siehe Tabelle 1). Diese Verfahren erfassen Fähigkeiten, Fertigkei-

Tabelle 1: Psychologische Test- und Fragebogenverfahren – ein Überblick mit Beispielen [3]

Leistungstests (Fähigkeiten/Fertigkeiten einer Person)	**Persönlichkeitstests** (Eigenschaften/Einstellungen einer Person)	
	Fragebogen	**Projektive Tests**
Entwicklungstests (z.B. Eggenberger Rechentest) Intelligenztests (z.B. Intelligenz-Struktur-Test) Allgemeine Leistungstests (z.B. d2-Aufmerksamkeits- und Konzentrationstest) Schultests (z.B. Skalen zur Erfassung der Lern- und Leistungsmotivation) Spezielle Funktionsprüfungs- und Eignungstests (z.B. Inventar sozialer Kompetenzen)	Einstellungsfragebogen (z.B. Test für Medizinische Studiengänge) Interessentests (z.B. Berufs-Interessentest) Klinische Tests (z.B. Beck-Depressions-Inventar) etc.	Formdeuteverfahren (z.B. Rohrschach-Test) Vebal-thematische Verfahren (z.B. Thematischer Apperzeptionstest) Zeichnerische Verfahren (z.B. Baum-Test) Gestaltungsverfahren (z.B. Thematischer Gestaltungstest)

ten, Eigenschaften und Zustände von Personen, die oftmals nicht direkt beobachtbar sind. Diese Merkmale werden mit Hilfe von beobachtbarem Verhalten erschlossen und als Konstrukte bezeichnet. Bekannte, in der medizinischen Ausbildungsforschung verwendete psychologische Konstrukte sind z.B. Motivation, Selbstwirksamkeit, Resilienz, Reflexionsfähigkeit, Empathie. Da Konstrukte nicht direkt gemessen werden können, werden sie als *latente Variablen* bezeichnet, für die Items als Indikatoren herangezogen werden [2].

Die meisten psychologischen Test- und Fragebogenverfahren beruhen auf den Annahmen und den Konstruktionsprinzipien der sog. *Klassischen Testtheorie (KTT)* [4], [5], mit der berücksichtigt wird, dass Messungen einzelner Personen über verschiedene Messungen hinweg variieren können. Das Grundkonzept beinhaltet die Annahme, dass der beobachtete Wert X einer Person in einem Test aus einem „wahren“ Wert (true score) der Person und einem zufälligen Messfehler (random measurement error) besteht. Das Ergebnis eines Intelligenztests würde dementsprechend zum einen von der tatsächlichen Intelligenz der untersuchten Person, zum anderen aber auch von unsystematischen Einflüssen, z.B. tageszeitlich bedingten Leistungsschwankungen beeinflusst. Würde man unendlich viele Messungen durchführen, dann entspräche der Mittelwert dieser Messungen dem tatsächlichen Intelligenzwert.

Neben der KTT gibt es die *Probabilistische Testtheorie* (PTT), die manchmal auch *Item-Response Theorie* (IRT) genannt wird. Diese Theorie beruht auf der Annahme, dass die Wahrscheinlichkeit einer bestimmten Antwort auf ein Item von Merkmalen des Items und der Ausprägung des zu messenden latenten Merkmals der Person abhängt [5]. Nach Rost [4] handelt es sich bei den beiden Testtheorien KTT und PTT nicht, wie häufig beschrieben, um konkurrierende, sondern um komplementäre Verfahren, da die eine Theorie dort ansetzt, wo die andere aufhört bzw. weil beide Testtheorien weitgehend auf denselben Annahmen beruhen. Nähere Ausführungen zur PTT finden sich in Bühner [2] und Döring und Bortz [5].

3. Wie werden Test- bzw. Fragebogenverfahren entwickelt?

Bei der Entwicklung von Test- bzw. Fragebogenverfahren sind sog. Testgütekriterien zu beachten (siehe Tabelle 2). Die Entwicklung solcher Verfahren beginnt mit der Festlegung bzw. Definition des zu erfassenden Konstrukts. Danach werden Items (Aufgaben oder Aussagen) formuliert und das Antwortformat ausgewählt. Nach einem Vortest wird das Verfahren an einer oder mehreren Stichproben gezielt erprobt. Wenn die dafür erforderliche Menge an Daten vorliegt, wird analysiert, wie zuverlässig der Test bzw. Fragebogen das Konstrukt misst (Reliabilität) und ob er das Konstrukt misst, das er zu messen beansprucht (Validität).

3.1. Definition des Konstrukts

Zur Definition und Operationalisierung des Konstrukts werden, soweit vorhanden, Theorien oder Modelle herangezogen. Beispiele für Konstrukte, zu denen es gute Theorien und Modelle gibt, auf deren Grundlage psychologische Tests entwickelt wurden, sind etwa Motivation und Lernstrategien. Sind Theorien und Modelle nicht verfügbar, kann der *Konstruktraum* nach ausgiebigem Literaturstudium eingeengt und es können Indikatoren (z.B. konkrete Aussagen oder Verhaltensweisen) des Konstrukts bestimmt werden. Ein aktuelles Beispiel aus der medizinischen Ausbildungsforschung, bei dem ein solches Vorgehen notwendig ist, ist die *Reflexionsfähigkeit*. Hier gibt es zwar auch verschiedene Modelle und Theorien, diese unterscheiden sich allerdings teilweise deutlich darin, was unter *Reflexionsfähigkeit* jeweils ver-

Tabelle 2: Traditionelle Test-Gütekriterien [2], [5], [10], [19]

Güte-kriterien[1]	Definition und Arten der Gütekriterien[1]	
Objektivität	Unter Objektivität versteht man den Grad, in dem die Ergebnisse eines Tests unabhängig von den Durchführungsbedingungen, der Auswertung und der Interpretation des Tests zustande kommen.	**Durchführungsobjektivität**: Die Durchführungsbedingungen eines Tests sollten immer gleich bzw. vergleichbar sein. **Auswertungsobjektivität**: Unabhängigkeit des ermittelten Testwertes von der Person, die den Test auswertet. **Interpretationsobjektivität**: Unabhängigkeit der Interpretation des ermittelten Testwerts von der Person, die die Interpretation vornimmt.
Reliabilität	Die Reliabilität eines Tests gibt den Grad der Messgenauigkeit eines Messwerts an, d.h. wie stark ein Messwert von zufälligen Messfehlern beeinflusst ist. Um die Reliabilität zu bestimmen, sind die nebenstehenden Verfahren möglich, die grundsätzlich auf dem Prinzip der mehrfachen Messung eines Merkmals beruhen.	**Retest-Reliabilität** (Merkmalsstabilität): Korrelation von Testleistungen, die zu zwei verschiedenen Zeitpunkten ermittelt wurden. **Paralleltest-Reliabilität** (Bedingungsstabilität): Korrelation der Messwerte von zwei Tests, die dieselbe Eigenschaft bzw. Fähigkeit mittels verschiedener Items erfassen. **Testhalbierungs-Reliabilität**: Korrelation der Messwerte der einen Testhälfte mit den Messwerten der zum gleichen Zeitpunkt erhobenen anderen Testhälfte. **Innere Konsistenz** (Messgenauigkeit zu einem bestimmten Messzeitpunkt): Korrelation jedes einzelnen Testitems mit jedem anderen, zum selben Zeitpunkt gemessenen Item korrigiert um die Testlänge. Einer der gebräuchlichsten Kennwerte der inneren Konsistenz ist Cronbachs α.
Validität	Die Validität gibt an, ob der Test das misst, was er zu messen beansprucht. Im Wesentlichen geht es darum überzeugend nachzuweisen, dass die Schlüsse, die aufgrund eines Testergebnisses im Hinblick auf das Verhalten oder die Eigenschaften einer Person gezogen werden, angemessen sind.	**Inhaltsvalidität** (Expertenbeurteilung): Die Items eines Tests stellen eine repräsentative Stichprobe des zu erfassenden Merkmalsbereichs dar. **Augenscheinvalidität** (Laienbeurteilung): Der Anspruch eines Tests, einen bestimmten Merkmalsbereich zu erfassen erscheint Laien bzw. Testpersonen vom bloßen Augenschein her gerechtfertigt. **Konstruktvalidität**: Der Test korreliert inhaltlich und theoretisch nachvollziehbar und hypothesenkonform mit anderen Konstrukten **Kriteriumsvalidität**: Der Test korreliert positiv mit inhaltlich korrespondierenden Merkmalen, die außerhalb der Testsituation gemessen werden können.
Ausgewählte Nebengütekriterien	**Vergleichbarkeit**: Wenn ein Test mehrmals von einer Person zu beantworten ist oder um das Abschreiben in Gruppentestungen zu erschweren, kann es vorteilhaft sein, einen Test in Parallelform bzw. Tests mit demselben Gültigkeitsbereich vorzulegen. **Ökonomie** eines Tests ist gegeben, wenn er, gemessen am diagnostischen Erkenntnisgewinn, relativ wenig Ressourcen wie Zeit, Geld etc. beansprucht. **Nützlichkeit**: Ein Test ist dann nützlich, wenn er etwas misst oder vorhersagt, für das ein praktisches Bedürfnis besteht. **Fairness** liegt vor, wenn Personen z. B. nach Alter, Geschlecht, Sprache die gleichen Chancen auf ein entsprechendes Testergebnis haben.	

[1] Veröffentlichungen u.a. von Loevinger [9] und Messick [13] legen vor allem für das Gütekriterium Validität eine andere Darstellungsweise mit einem für uns angemesseneren Denkmodell nahe. In diesem Artikel schließen wir uns jedoch Döring und Bortz [5] an, die sich zur besseren Vermittlung der Testgütekriterien für eine *Strategie der bewusst ungenauen Sprechweise* entschieden haben.

standen wird. Insofern müsste für die Entwicklung eines Testverfahrens für *Reflexionsfähigkeit* zunächst definiert werden, welche Indikatoren von *Reflexionsfähigkeit* auf Grundlage welcher Vorarbeiten berücksichtigt werden sollen. Im Rahmen der Konstrukt-Definition soll zudem auch ermittelt werden, inwieweit Beziehungen sowie Überschneidungen bzw. Überlappungen zu anderen Konstrukten bestehen (nomologisches Netzwerk) [2]. So wird z.B. diskutiert, inwiefern die Persönlichkeitseigenschaft *Offenheit für Erfahrungen* mit Kreativität in Verbindung steht [6].

Die Güte der Definition des Konstrukts entscheidet darüber, wie leicht sich Items generieren lassen. Eine detaillierte Definition, die erforderliche Abgrenzungen gegenüber anderen Konstrukten berücksichtigt, erhöht darüber hinaus die Wahrscheinlichkeit für die *Inhaltsvalidität* des Konstrukts [1], [2].

3.2. Generierung und Formulierung von Items

Bei der Generierung von Items kann auf unterschiedliche Quellen zurückgegriffen werden [1]. So können Items

- aus *Theorien* bzw. nach einem ausgiebigen *Literaturstudium* bzw. nach einer systematischen Literaturrecherche abgeleitet werden,
- aus Ergebnissen von *Voruntersuchungen* (Interviews, Fokusgruppen-Gespräche etc.) generiert werden,
- in Anlehnung an bestehende *Testverfahren* formuliert werden,
- von *Expertinnen und Experten* formuliert werden.

Auch bei der Item-Generierung sind die Ziele des zu konstruierenden Tests zu berücksichtigen [2]. Ist das Ziel, Eigenschafts- oder Fähigkeitsausprägungen von Personen zu erfassen, so sollten inhaltsvalide Items formuliert werden. Ein Test zur Erfassung von *Progredienzangst*, d.h. Angst vor dem Voranschreiten einer Erkrankung, ist dann inhaltsvalide, wenn die Testitems eine repräsentative Stichprobe des gesamten Bereichs von *Progredienzangs*t darstellen (z.B. kognitive, emotionale und verhaltensbezogene Aspekte). Es sollte dabei darauf geachtet werden, dass mit den Items nur ein Konstrukt erfasst wird. Darüber hinaus sollten alle Indikatoren eines Konstrukts miteinander korrelieren [2].
Zur Sicherung der *Inhaltsvalidität* des Tests sollte bei der Item-Generierung auf eine repräsentative und ausreichende Item-Menge geachtet werden. Die Anzahl der Items des Testentwurfs sollte größer sein als die geplante Item-Anzahl der Endversion [2].
Vor Beginn der Itemformulierung sollte darüber entschieden werden, wie die Items formuliert werden sollen. Sie können beispielsweise wie folgt formuliert werden:

- *In Frageform*: Fühlen Sie sich von Angehörigen anderer Gesundheitsberufe respektiert?
- *Als Statements:* Ich fühle mich von Angehörigen anderer Gesundheitsberufe respektiert.
- *In 1. Person Singular:* Ich arbeite gerne mit Angehörigen anderer Gesundheitsberufe zusammen.
- *In unpersönlicher Form:* Man arbeitet hier gerne mit Angehörigen anderer Gesundheitsberufe zusammen.

Die Items sollten inhaltlich schlüssig und verständlich sein [1], [2]. Unter anderem bedeutet dies, dass Fremdwörter oder eine komplizierte Satzkonstruktion zu vermeiden sind. Auch sollten die Items inhaltlich eindeutig sein. Hierzu sind nach Möglichkeit u.a. Konditionalaussagen oder Konjunktionen zu vermeiden. Auch sollten Negationen (insbes. doppelte Verneinungen) vermieden werden.

3.3. Auswahl des Antwortformats

Genauso wichtig wie die Formulierung der Items ist die Auswahl passender Antwortvorgaben. Häufig finden bei psychologischen Test- und Fragebogenverfahren Ratingskalen (meist sog. Likert-Skalen) Anwendung, deren Kategorien bzw. Abstufungen unterschiedlich benannt werden. Oft anzutreffen sind Benennungen wie „trifft nicht zu“ bis „trifft zu“ oder „sehr schlecht“ bis „sehr gut“. Ratingskalen können zudem unterschiedlich abgestuft sein. Hierbei sind Antwortskalen mit bis zu 7 Stufen akzeptabel [2]. Des Weiteren ist zu klären, ob die Antwortstufen der Items unipolar (z.B. „nie“ bis „sehr oft“) oder bipolar (z.B. „Ablehnung „teilweise Ablehnung“, „weder Ablehnung noch Zustimmung“, „teilweise Zustimmung“, „Zustimmung“) vorgegeben werden sollen. Neben der verbalen Benennung der Antwortstufen können auch visuelle Hilfsmittel verwendet werden (z.B. Smileys).

3.4. Formulierung der Instruktion

Die Instruktion hat zum Ziel, Befragte mit dem Inhalt und Ziel der Befragung vertraut zu machen, Hinweise zur Beantwortung des Fragebogens zu geben und über datenschutzrechtliche Regelungen aufzuklären [7]. Sie hat eine zentrale Funktion, denn sie bereitet nicht nur auf die Beantwortung des Tests vor, sondern kann bei den zu befragenden Personen eine Vor-Einstellung in Bezug auf die zu erledigende Aufgabe erzeugen [1]. Eine Instruktion wird meist erst am Ende des Konstruktionsprozesses formuliert, wenn die Items und Antwortalternativen festgelegt sind. Neben einer Angabe des Ziels bzw. des Zwecks des Tests bzw. Fragebogens enthält eine Instruktion i.d.R. Hinweise, dass

- die Teilnahme freiwillig ist und keine Nachteile bei einer Nichtteilnahme zu befürchten sind,
- alle Items zu lesen und zügig zu beantworten sind,
- die Items nacheinander zu bearbeiten sind und kein Item ausgelassen werden soll, auch wenn dies einmal schwierig erscheinen sollte, und dass in diesem Fall stets angekreuzt werden sollte, was „am ehesten“ zutrifft,
- die Anonymität bzw. die vertrauliche Behandlung der individuellen Angaben entsprechend den geltenden datenschutzrechtlichen Bestimmungen gewährleistet werden.

3.5. Durchführung von Vortests

Die Durchführung eines oder auch mehrerer Vortests ist eine weitere wichtige Voraussetzung der Entwicklung eines Test- oder Fragebogenverfahrens. Für dessen Durchführung existieren jedoch keine allgemein akzeptierten Regeln. Beispielsweise variieren die Angaben sehr stark, wenn es darum geht, die Höhe der hierfür notwendigen Fallzahlen festzulegen [8]. In der Regel wird jedoch eine kleine Zahl von Personen aufgefordert, Rückmeldung über die Verständlichkeit der Items und der Instruktion zu geben und über Schwierigkeiten zu berichten, die bei der Bearbeitung des Verfahrens aufgefallen sind. Wichtig ist, dass diese Personen der späteren Zielgruppe des Tests bzw. Fragebogens möglichst ähnlich sind, z.B. was das Sprachverständnis angeht. Vortests liefern auch Informationen über die benötigte Durchführungszeit, das Interesse der Befragten an der Thematik sowie über die

Häufigkeitsverteilungen der Antworten. Auf Basis der Rückmeldungen wird das Verfahren dann ggf. modifiziert.

4. Teststatistische Überprüfung von psychologischen Test- und Fragebogenverfahren

In Anlehnung an die von Loevinger [9] herausgearbeiteten Phasen kann der Prozess der teststatischen Überprüfung wie folgt eingeteilt werden:

- *Inhaltliche Phase* (substantive phase): Während dieser Phase wird das Messinstrument theoretisch und unter Einbeziehung verfügbarer Literatur fundiert. Es werden Vortests durchgeführt, um die Verständlichkeit der Items und Probleme bei deren Beantwortung abzuklären.
- *Strukturelle Phase* (structural phase): Das Hauptaugenmerk dieser zweiten Phase richtet sich auf die Überprüfung der strukturellen (z.B. faktorielle Struktur) und weiterer psychometrischen Eigenschaften (z.B. Item-Korrelationen) des Verfahrens.
- *Externe Phase* (external phase): In dieser Phase sollte das Ausmaß der Übereinstimmung des Messinstruments mit anderen Kriterien und ggf. ähnlichen Verfahren überprüft werden.

Alle bisherigen Beschreibungen in diesem *Gewusst-wie-Artikel* lassen sich der *inhaltlichen Phase* zuordnen (siehe Tabelle 3). In den nachfolgenden Abschnitten geht es schwerpunktmäßig um die konkrete teststatistische Überprüfung der Test- bzw. Fragebogenverfahren, die den anderen beiden Phasen zuzuordnen sind.

4.1. Strukturelle Phase

In der *inhaltlichen Phase* kann bereits die *Augenschein- und Inhaltsvalidität* eines Test- bzw. Fragebogenverfahrens sichergestellt werden. Die Überprüfung der strukturellen und psychometrischen Eigenschaften von Tests bzw. Fragebögen kann jedoch erst dann stattfinden, wenn das Verfahren von Personen der jeweiligen Zielgruppe beantwortet wurde (Datenerhebung). Es sollte zunächst eine Dimensionsanalyse mittels Faktorenanalysen (statistische Verfahren, die die Variablen gemäß ihrer Interkorrelation bündeln) erfolgen (*faktorielle Validität*) und anschließend eine Bestimmung der Reliabilität des Tests und eine Itemanalyse durchgeführt werden [10]. Ist der Stichprobenumfang jedoch zu gering [2], um Dimensionsanalysen durchzuführen, können zunächst vorläufige Reliabilitätsberechnungen durchgeführt werden und die Items im Hinblick auf ihre Schwierigkeit, Trennschärfe und Interkorrelationen (Itemanalysen) analysiert werden (siehe Tabelle 4).
Die in der einschlägigen Literatur angegebenen erforderlichen Stichprobengrößen zur Berechnung von Faktorenanalysen variieren sehr stark. Nach MacCallum et al. [11] sind die gängigen Faustregeln zur Planung der Stichprobengröße problematisch, da diese von der Anzahl der Items pro Faktor und der Höhe der Kommunalität (Anteil der Varianz einer Variablen, der durch die Faktoren erklärt wird) eines jeden Items bestimmt wird. Die Kommunalitäten sind jedoch in der Regel nicht vorab bekannt. Von daher soll hier trotz der genannten Problematik zur groben Orientierung zumindest erwähnt werden, dass für Faktorenanalysen verschiedentlich empfohlen wurde, eine Anzahl von zu Befragenden einzuplanen, die mindestens fünf- bis zehnmal so groß ist wie die Anzahl der Items.
Ist die Stichprobengröße ausreichend zur Durchführung von Faktorenanalysen und liegt eine Hypothese bzw. ein Modell zu den Dimensionen des Tests vor, sollte eine *konfirmatorische Faktorenanalyse* durchgeführt werden. Gibt es keine gesicherten Annahmen über die Zusammenhänge zwischen den Items, ist eine *exploratorische Faktorenanalyse* zu empfehlen.

4.2. Externe Phase

Die Validierung eines Test- bzw. Fragebogenverfahrens ist theoretisch und methodisch anspruchsvoll und sollte nie als abgeschlossen betrachtet werden [5], [12]. Insofern sollte strenggenommen auch nicht davon gesprochen werden, dass ein Test bzw. Fragebogen valide ist, da Validität keine Eigenschaft von Tests bzw. Fragebögen ist (siehe 4.2.2). Mehr oder weniger valide können nur Aussagen und Schlussfolgerungen sein, die auf der Grundlage von Test- bzw. Fragebogenergebnissen getroffen werden. Die Validierung von Test- bzw. Fragebogenverfahren (bzw. genauer von Test- bzw. Fragbogenergebnissen) beinhaltet verschiedene Aspekte. Dabei hat sich das Verständnis, welche Indikatoren als Hinweise auf Validität gelten können, im Lauf der Zeit verändert. Nachfolgend wird zunächst das traditionelle Validitätskonzept dargestellt. Im Anschluss daran wird der Validitätsansatz von Messick [13] beschrieben, der den traditionellen Ansatz ergänzt.

4.2.1. Konstrukt- und Kriteriumsvalidität

Zunächst kann festgestellt werden, ob das im Test- bzw. Fragebogenverfahren erfasste Konstrukt inhaltlich und theoretisch begründet mit anderen *theoretischen Konstrukten* korreliert (*Konstruktvalidität*) und/oder ob die Test- bzw. Fragebogenwerte positiv mit inhaltlich korrespondierenden manifesten Merkmalen außerhalb der Testsituation im Zusammenhang stehen (*Kriteriumsvalidität*) [5].
Zur Feststellung der *Konstruktvalidität* können Messinstrumente eingesetzt werden, die entweder sog. konstruktnahe oder konstruktferne Merkmale erfassen. Im ersten Fall würde nach Campbell und Fiske [14], die *konvergente Validität* überprüft, im zweiten Fall die *diskriminante Validität*. Zur *Konstruktvalidität* zählt ebenfalls die zuvor beschriebene *faktorielle Validität* (siehe 4.1). Auch besteht die Möglichkeit, Unterschiede in den Testwerten ausgewählter Gruppen zu analysieren. D.h. ausgehend von theoretischen Überlegungen werden Unterschiede in den Testwerten verschiedener Gruppen postuliert (z.B.

Tabelle 3: Übersicht der Phasen der Konstruktion von Test- und Fragebogenverfahren (siehe Kapitel 4)

Phasen	Ablauf	Relevante Aspekte	Gütekriterien
Inhaltliche Phase (substantive phase)	Festlegung und Definition des Konstrukts	– Theorien, Modelle – Literaturrecherche – Nomologische Netzwerke	
	Generierung und Formulierung von Items	Generierung Theorien, Literatur, Voruntersuchungen (z.B. Fokusgruppen), bestehende Tests, Experten Formulierung Frageform, Aussagen, 1. Person Singular, unpersönliche Form, keine komplizierte Satzform, keine Fremdwörter, keine doppelte Verneinung etc.	Augenscheinvalidität (face validity) Inhaltsvalidität (content validity)
	Auswahl des Antwortformats	– Ratingskalen (meist Likert) – Gestaltung der Ratingskalen: unipolar, bipolar, visuelle Hilfsmittel (z.B. Smileys)	
	Formulierung der Instruktion	– Information über Ziel/Zweck des Tests/Befragung – Hinweis alle Items zu lesen und zügig zu beantworten – Hinweis Items nacheinander zu beantworten, ohne ein Item auszulassen – Zusicherung der Anonymität, Hinweise auf Datenschutz – etc.	
	Durchführung von Vortests	– Rückmeldung zur Verständlichkeit der Items, der Instruktion – Rückmeldung zu Schwierigkeiten bei der Beantwortung – Erfassung der benötigten Durchführungszeit – etc.	
Strukturelle Phase (structural phase)	Überprüfen der Teststruktur	– Überprüfung der Dimensionalität des Tests mittels konfirmatorische bzw. exploratorische Faktorenanalysen – Berechnung der Reliabilität – Durchführung von Itemanalysen (M, SD, Schwierigkeitsindex, Trennschärfen)	Faktorielle Validität Reliabilität
Externe Phase (external phase)	Sammeln von Evidenz für die Gültigkeit von Testergebnissen	– Überprüfung von Konstrukt- und Kriteriumsvalidität – Überprüfen von Hypothesen und Argumenten, die auf Grundlage der Testergebnisse formuliert werden	Validitätsarten, z. B.: - konvergente - diskriminante - retrospektive - konkurrente - Gruppenunterschiede

Alter, sozioökonomischer Status, Schulbildung) und empirisch überprüft [10]. Sofern sich diese Unterschiede bestätigen, wird dies als Beleg der Validität interpretiert. Bei der *Kriteriumsvalidität* lassen sich ausgehend vom Zeitpunkt der Erfassung des Außenkriteriums mehrere Arten von Validität unterscheiden [5]. Die *retrospektive Validität* wird überprüft, wenn Werte eines Kriteriums (z.B. zurückliegende Schulnoten) zeitlich vor dem Einsatz des zu validierenden Tests (z.B. Schulleistungstest) erhoben wurden. Bei der *konkurrenten Validität*, auch *Übereinstimmungsvalidität* genannt, werden die Werte des Kriteriums (z.B. die in ärztlichen Konsultationen angegebene Beschwerden wie Schlaf- und Lustlosigkeit) (fast) zum selben Messzeitpunkt erfasst wie die zu validierenden Testwerte (z.B. Ergebnisse eines Messinstruments zur Erfassung der Ausprägung von Depression). Bei der

Tabelle 4: Beschreibung von teststatistischen Kennwerten

Kennwerte	Beschreibung
Mittelwert	Der Durchschnittswert bzw. das arithmetische Mittel ist die Summe aller Werte einer Messreihe geteilt durch die Anzahl der Werte.
Standardabweichung	Die Standardabweichung ist ein Maß für die Streubreite der Werte eines Merkmals um deren Mittelwert.
Schwierigkeitsindex	Der Schwierigkeitsindex ist ein Maß dafür, wie schwer oder leicht eine Aufgabe von Befragten einer Stichprobe zu beantworten bzw. zu lösen ist.
Trennschärfe	Die Trennschärfe informiert darüber, wie gut ein Item zwischen Personen mit niedriger und hoher Merkmalsausprägung trennt.
Kommunalität eines Items	Der Anteil an der Gesamtvarianz einer Variablen, der durch alle Faktoren gemeinsam erklärt werden kann.

prognostischen Validität wird der Kriteriumswert (z.B. Studienleistung) *später* als der zu validierende Testwert (z.B. Ergebnisse eines Eignungstests zum Medizinstudium) erhoben. Die Bestimmung der *Kriteriumsvalidität* setzt voraus, dass das gewählte Außenkriterium reliabel und valide ist.

Die *inkrementelle Validität* zählt ebenfalls zur *Kriteriumsvalidität*, wird jedoch eher selten überprüft. Wenn doch eine Überprüfung erfolgt, wird ein herkömmliches Verfahren herangezogen, das das Gleiche zu messen beansprucht wie das zu validierende Verfahren. Dabei sollte das neue Verfahren die Vorhersage des Außenkriteriums signifikant verbessern [5].

4.2.2. Argumentationsbasierte Validierungskonzepte

Das im vorigen Abschnitt beschriebene klassische Validitätskonzept wurde von Messick [13] erweitert. Die von ihm beschriebenen sechs generellen Validitätsaspekte, die für alle diagnostischen Messungen im Bildungsbereich gelten, basieren auf der Grundidee, dass die Validität einer diagnostischen Messung nicht allein als numerischer Koeffizient zu betrachten ist, sondern als theoretisch und empirisch fundiertes Argument für die Gültigkeit von Testwertinterpretationen. M.a.W. „It is incorrect to use the unqualified phrase *the validity of the test*" ([15], S.11), denn die Ergebnisse sind nicht nur eine Funktion der Items, sondern auch abhängig von den antwortenden Personen und dem Kontext der Bewertung [13]. Validität kann demnach als Argument für die Gültigkeit der Interpretation von Testwerten auf Grundlage von Evidenzen bzw. Erkenntnissen bezüglich dieser sechs Aspekte verstanden werden. In Tabelle 5 sind die von Messick beschriebenen Validitätsaspekte dargestellt. Dabei wird ersichtlich, dass nur die Aspekte substanzielle Validität, Generalisierbarkeit und Konsequenzen den traditionellen Ansatz ergänzen (siehe Tabelle 5).

Ergänzend hat Messick [13] auf zwei mögliche Störfaktoren hingewiesen, die die Validität beeinträchtigen können. Das Konstrukt kann *unterrepräsentiert* sein, indem es zu eng gefasst und wichtige Dimensionen oder Facetten des Konstrukts nicht berücksichtigt wurden. Das wäre beispielsweise dann der Fall, wenn ein Test für Leistungsangst nur die emotionale Komponente erfasst und die kognitiven und physiologischen Komponenten außer Acht lässt. Die Validität kann aber auch durch *konstrukt-irrelevante Varianz* eingeschränkt werden, wenn Testaufgaben beispielsweise für einige Personen zu schwer oder zu leicht sind [13]. Dies ist z.B. der Fall, wenn die korrekte Beantwortung von Aufgaben in einem Mathematiktest auch von unangemessen hohen Anforderungen an das Sprachverständnis der antwortenden Personen abhängt. Diese Erweiterungen des klassischen Validitätskonzepts werden mittlerweile u.a. von der *American Educational Research Association* (AERA) und der *American Psychological Association* (APA) vertreten [15], [16].

5. Übersetzung eines Test- oder Fragebogenverfahrens

Die Übersetzung von Tests bzw. Fragebögen erfolgte in der Vergangenheit häufig mit der Methode der Rückübersetzung. D.h. zuerst wurde das Verfahren in die Zielsprache übersetzt, dann wurde diese Übersetzung von einer anderen Person zurückübersetzt [17] und anschließend wurden die ursprüngliche und die rückübersetzte Version miteinander verglichen. Eine einfache Rückübersetzung kann jedoch nicht alle Übersetzungsprobleme beseitigen, daher werden mittlerweile mehrstufige Übersetzungsprozesse empfohlen [17]. Beispielsweise wird gemäß den Richtlinien des *European Social Survey Programme* zur Übersetzung von Fragebögen unter dem Akronym *TRAPD* ein fünfstufiger Übersetzungsprozess vorgeschlagen: *T*ranslation, *R*eview, *A*djudication (deciding on a version), *P*re-testing und *D*ocumentation [18]. Diese Schritte sollten von Beginn an in Teamarbeit erfolgen. Auch bei der Übersetzung eines Tests ist eine vollständige statistische Überprüfung der übersetzten Version erforderlich.

Tabelle 5: Validitätsansatz von Messick [5], [13], [20]

Validitätsaspekt	Bedeutung
Inhaltlich (content)	Nachweis für die inhaltliche Relevanz, Repräsentativität und Qualität des erfassten Konstrukts.
Substanziell (substantive)	Nicht nur die Inhalte eines diagnostischen Verfahrens sollten repräsentativ für den zu erfassenden Merkmalsbereich sein, sondern auch die kognitiven Prozesse, die zur Beantwortung der Testaufgaben benötigt werden. D.h., die Probanden sollten sich primär mit den Inhalten des Tests auseinandersetzen (und nicht etwa damit, eine komplizierte Formulierung der Testinstruktion oder eines Items zu verstehen). Die Übereinstimmung zwischen den in der Testsituation ablaufenden kognitiven Prozessen und denjenigen, die als zentral für das zu testende Konstrukt angenommen werden, sollte empirisch überprüfbar sein. Dies kann z.B. während der Pilotierung des Tests in der Vortestphase u.a. mit Hilfe der „Methode des lauten Denkens“ oder mit der Analyse von Augenbewegungen erfolgen.
Strukturell (structural)	Übereinstimmung von theoretischem Modell bzw. Strukturen des Konstrukts und dem empirischen Messmodell: Die empirisch z.B. mittels Faktorenanalyse ermittelte Struktur der Testdaten sollte mit den Annahmen über die Struktur des Konstrukts übereinstimmen (z.B. Aspekte von Motivation, Intelligenz etc.).
Generalisierbarkeit (generalizability)	Berücksichtigt, inwieweit sich die Eigenschaften und Interpretationen der Testwerte auf andere Aufgaben, Bevölkerungsgruppen, Settings verallgemeinern lassen.
External (external)	Bezieht sich auf hypothesenkonforme Zusammenhänge zwischen Testergebnisse mit anderen Merkmalen (konvergente, diskriminante Validität etc.)
Konsequenzen (consequential)	Berücksichtigt positive wie negative Folgen der Verwendung von Testergebnissen. *Teaching-to-the-Test* wäre ein Beispiel für eine nicht erwünschte Folge des Einsatzes eines Schulleistungstests. Der Unterricht wird darauf umgestellt, zielgerichtet auf den bevorstehenden Test vorzubereiten.

6. Fazit

Die Konstruktion von Tests bzw. Fragebögen setzt ein gut definiertes Konstrukt oder zumindest einen konkret beschriebenen Konstruktraum voraus. Auf dieser Grundlage können Items formuliert werden, die inhaltlich valide und gut verständlich sein müssen und sich an den Zielen des Verfahrens orientieren. Hat sich das Verfahren in Vortests an kleinen Personengruppen bewährt, können seine strukturellen (Dimensionalität) und weiteren psychometrischen (Reliabilität, Validität etc.) Eigenschaften anhand von umfangreicheren Datenerhebungen überprüft werden. Um die Validität der Testergebnisse zu bestimmen, müssen verschiedene Aspekte berücksichtigt werden. Diese beziehen sich vor allem auf das zu messende Konstrukt und seine theoretische Einbettung sowie seine Beziehung zu anderen Variablen, aber auch auf den Kontext der Messung und die Konsequenzen, die aus den Testergebnissen abgeleitet werden.

ORCIDs der Autorin und des Autors

- Marianne Giesler: [0000-0001-9384-2343]
- Götz Fabry: [0000-0002-5393-606X]

Interessenkonflikt

Die Autorin und der Autor erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

Literatur

1. Mummendey HD, Grau I. Die Fragebogen-Methode. 5. Aufl. Göttingen: Hogrefe Verlag; 2008.
2. Bühner M. Einführung in die Test- und Fragebogenkonstruktion. 3. aktual. u. erw. Aufl. München: Pearson Studium; 2011.
3. Brähler E, Holling H, Leutner D, Petermann F. Brickencamp Handbuch psychologischer und pädagogischer Tests. 3. Aufl. Göttingen: Hogrefe; 2002.
4. Rost J. Was ist aus dem Rasch-Modell geworden? Psych Rundsch. 1999;50(3):140-156. DOI: 10.1026//0033-3042.50.3.140
5. Döring N, Bortz J. Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5. vollst. überarb., aktual. u. erw. Aufl. Berlin: Springer; 2016. DOI: 10.1007/978-3-642-41089-5
6. King LA, Walker LM, Broyles SJ. Creativity and the five-factor model. J Res Pers. 1996;30(2):189-203.
7. Reinders H. Fragebogen. In: Reinders H, Ditton H, Gräsel C, Gniewosz B, editors. Empirische Bildungsforschung. Strukturen und Methoden. Wiesbaden: VS Verlag für Sozialwissenschaften; 2011. p.53-65. DOI: 10.1007/978-3-531-93015-2_4
8. Porst R. Im Vorfeld der Befragung: Planung, Fragebogenentwicklung, Pretesting. ZUMA-Arbeitsbericht, 1998/02. Mannheim: Zentrum für Methoden und Analysen (ZUMA); 1998. URN: urn:nbn:de:0168-ssoar-200484
9. Loevinger J. Objective tests as instruments of psychological theory. Psychol Rep. 1957;3(3):635-694. DOI: 10.2466/pr0.1957.3.3.635
10. Lienert GA. Testaufbau und Testanalyse. 2. durchges. u. verb. Aufl. Weinheim: Beltz; 1961.
11. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample Size in Factor Analysis. Psychol Method. 1999;4(1):84-99. DOI: 10.1037/1082-989X.4.1.84

12. Repke L, Birkenmaier L, Lechner CM. Validity in Survey Research - From Research Design to Measurement Instruments. Mannheim: GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines); 2024. DOI: 10.15465/gesis-sg_en_048

13. Messick S. Validity of Psychological Assessment. Validation of Inferences from Persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741-749. DOI: 10.1002/j.2333-8504.1994.tb01618.x

14. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56(2):81-105.

15. AERA, APA, NCME. Standards for Educational and Psychological Testing. Washington: American Psychological Association (APA); 2014.

16. Schaper N. Validitätsaspekte von Kompetenzmodellen und -tests für hochschulische Kompetenzdomänen. In: Musekamp F, Spöttl G, editors. Kompetenz im Studium und in der Arbeitswelt. Nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen. Frankfurt, M: Lang; 2014. p.21-48.

17. Su CT, Parham LD. Generating a valid questionnaire translation for cross-cultural use. Am J Occup Ther. 2002;56(5):581-585. DOI: 10.5014/ajot.56.5.581

18. European Social Survey. ESS Round 11 Translation Guidelines. London: ESS ERIC Headquarters; 2022. Zugänglich unter/available from: https://www.europeansocialsurvey.org/sites/default/files/2024-08/ESS_R11_Translation_Guidelines.pdf

19. Moosbrugger H, Kelava A. Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger H, Kelava A, editors. Testtheorie und Fragebogenkonstruktion. 3. vollst. neu bearb., erw. u. akt. Aufl. Heidelberg: Springer; 2020. p.7-26. DOI: 10.1007/978-3-540-71635-8_2

20. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-837. DOI: 10.1046/j.1365-2923.2003.01594.x

Korrespondenzadresse:
Marianne Giesler
Freiburg i. Brsg., Deutschland
Dr_M_Giesler@t-online.de

Bitte zitieren als
Giesler M, Fabry G. How do I develop a psychological test or questionnaire? GMS J Med Educ. 2026;43(1):Doc9. DOI: 10.3205/zma001803, URN: urn:nbn:de:0183-zma0018034

Artikel online frei zugänglich unter
https://doi.org/10.3205/zma001803

Eingereicht: *20.01.2025*
Überarbeitet: *13.05.2025*
Angenommen: *28.07.2025*
Veröffentlicht: *15.01.2026*