Methodologies of PD Assessment: Scales

Roman Görtelmeyer

Abstract

Scales and behavioral tests are the primary source of data in human research for measuring, modeling, and testing perception, thoughts, feelings, complains, and performance. In drug development scales and tests are used in clinical trials as tools to support diagnosis; evaluate states and traits, as well as lifestyle and performance of patients; or collect ratings by relatives and health carers before, during, and after treatment. Clinical scales can be classified according to several criteria; one classification is according to the intended responder and the method of data generation (FDA 2016; Walton et al. Value Health 18:741–752, 2015). Well-developed and ready-for-purpose scales and tests can be used to describe, characterize, or quantify (localize) objects and processes within a framework of a pharmacological clinical study or preferably of a drug development program of interest. In the present article the conceptualization, development, selection, analysis, and application of items and scales in various phases of clinical drug development are presented in an overview and discussed.

R. Görtelmeyer (✉)
Scientific Consulting, Frankfurt am Main, Germany

Biological and Clinical Psychology, University of Mannheim, Mannheim, Germany
e-mail: roman.goertelmeyer@t-online.de; goertelm@mail.uni-mannheim.de

Contents

Purpose and Rational 1
Procedure 5
Evaluation 12
Critical Assessment of the Method 13
Modification of the Method 20

References and Further Reading 21

Purpose and Rational

Scales and behavioral tests are the primary source of data in human research for measuring, modeling, and testing perception, thoughts, feelings, complains, and performance. In drug development scales and tests (for convenience further subsumed as clinical scale) are used in clinical trials as tools to support diagnosis; evaluate states and traits, as well as lifestyle and performance of patients; and collect ratings by relatives and health carers before, during, and after treatment or for the duration of a clinical study. Clinical scales have been classified according various criteria; one of these classifications is the qualification according to the intended responder and the method of data generation (FDA 2016; https://www.fda.gov/Drugs/%20DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm; Walton et al. 2015). As such "outcome assessments include survival, clinical outcome assessments (COAs), and biomarkers. A COA is any evaluation that can be influenced by human choices, judgment, or motivation. There are four types of COAs: patient-reported outcome

F.J. Hock, M.R. Gralinski (eds.), *Drug Discovery and Evaluation: Methods in Clinical Pharmacology*,
https://doi.org/10.1007/978-3-319-56637-5_27-1

(PRO), clinician-reported outcome (ClinRO), observer-reported outcome (ObsRO), and performance outcome (PerfO)" (Powers et al. 2017). Beyond these qualifications scales may be of great help in exploring the mechanisms of action on human behavior and attitudes. The assessment program of pharmacological intervention comprises evaluation of concept, methods, disease model, definition of (target) population, concept of effect, domains of potential drug response, and benefit and risk of pharmacological intervention and could include estimates of probable success in the market. In most phases of drug development, scales can be efficiently used for differential benefit of the program. Measurement and scaling are fundamental processes in the empirical sciences and of special importance in drug research and development. In natural science measurement may be merely regarded as the use of an existing scale to quantify the object, and scaling can be regarded as "... the assignment of objects to numbers according to a rule" (Stevens 1951). Measurement is only possible if some scale is defined. The use of "scale" in literature is not unique. The term is used in various contexts with different meanings. A more generic definition can be found in dictionaries, e.g., relative size or extent (Oxford English Dictionary).

In the present paper, scale is used synonymously for a standardized and validated questionnaire or behavioral test which delivers a score that has been empirically proven to measure or indicate the underlying characteristic or process of interest. The result of measurement by clinical scales usually is a score on an abstract dimension, e.g., probability, intensity, frequency, as well as change or difference, or the result of scoring and scaling will be transposed into some classification. The way many psychometricians are defining *scale* includes the concept, addressed by some task, a verbal expression, or a question, combined with a concept and design of response option, which may be some kind of differential verbalization or any other ordering of response format that is presented to the respondent. Further, a method for proper quantification of the response and some evidence on reliability and validity is expected. Scale defined in this sense is the standardized quantification of a response of someone in a well-defined test situation (e.g., Lienert and Raatz 1998). The definition of exact standards of measurement is related to units that refer to specific conditions and quantitative attributes. In the natural sciences, the metric international system of measurements (système international d'unités, SI) is used, in which scientifically spoken some quantities are designated as the fundamental units. The first fundamental units, referring to specific empirical conditions and quantitative attributes, were:

Meter (m) SI unit of length
Second (s) SI unit of time
Kilogram (kg) SI unit of mass
Kelvin (K) SI unit of temperature

In 1971, the last of today's accepted seven basic units was Mol (mol), the SI unit of amount of substance from which all other needed units can be derived.

With increasing use and relevance of electronic communication, the *Unified Code for Units of Measure* (UCUM) has gained fundamental importance. UCUM is "a code system intended to include *all* units of measures being contemporarily used in international science, engineering, and business... to facilitate unambiguous electronic communication of quantities together with their units. The focus is on electronic communication, as opposed to communication between humans." (Schadow and McDonald 2014, http://unitsofmeasure.org/ucum.html). The failure of the NASA Mars Climate Orbiter, which was destroyed on a mission to the Mars in September 1999 instead of entering the planet's orbit, was due to miscommunications about the value of forces (ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf.).

Social, educational, and clinical sciences lack standard measures and units. Continuous efforts are made by various working groups to develop frameworks and models to link physiological, pharmacological, and biological units and systems with behavior to understand basic dimensions of functioning. A prominent example is the Research Domain Criteria (RDoC) which is

"centered around dimensional psychological constructs (or concepts) that are relevant to human behavior and mental disorders, as measured using multiple methodologies and as studied within the essential contexts of developmental trajectories and environmental influences." (National Institute of Mental Health, NIH, https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml). This approach, like similar others, could be addressed in conceptual frameworks for drug development programs, not limited to neuropsychopharmacology. Ideally, in drug research and development, the measurement concept should be developed right from the beginning of the search for a new substance or method when first consideration on medical indications is in play or in case of planned modification of a known drug (FDA Clinical Outcome Assessment Qualification Program 2017). The early decision on clinical measures may help to link pharmacological evidence to the assessment of patient's benefit after approval by authorized bodies.

The measurement of hypothesized effect as well as the confirmation of known effects of a substance or compound needs a theory-based quantification strategy to make the hypothesis testable and enable new hypotheses. There are two major strategic mistakes that should be avoided from the perspective of methodology: usage of measurement tools of unknown validity or reliability in explorative or even pivotal clinical trials and usage of well-known measurement instruments in the wrong experimental setting or population.

Clinical measurement and scaling methods are used to describe, qualify, and quantify (localize) objects, constructs, or behavior during the clinical studies and document the treatment outcomes according to the study protocol. Some scales may either be used as diagnostic tools for the classification or severity grading of patients, and as criteria for selection of patients, or as outcome variables. In justified cases, scales may be used for both diagnostic purposes and outcome variables. In case of diagnostic tools, it is important that the scale has been shown to be of sufficient content and construct validity to cover essential aspects of the target disease, condition, symptom, or syndrome. Of special interest in binary decision are the sensitivity and specificity of the diagnostic instrument to support classification (e.g., "patient shows signs of disease X": yes/no), with "sensitivity" defined as the detection rate of correct positive and "specificity" defined as the detection rate of correct negative identified persons, e.g., not having symptoms of disease X. These criteria are prerequisites for the correct selection of valid samples from the target population as defined in the clinical drug development plan.

The observer-rated scales on the other side are helpful to collect additional data on treatment effects outside the visits with the doctor or investigator. They should always be introduced with an explicit recommendations on who (expert, physician, medicinal personal, etc.) and at which time during the study they have to be used. It is mandatory to define the accountable respondent in the study protocol to make sure that the appointed rater(s) will be trained for the correct use of the scale throughout the study.

Diagnosis for sure is a complex cognitive process which frequently, especially in very experienced experts, is emerging from implicit processing of information. Various studies have shown that medical experts are better in differential diagnostic, tend to process less information, and decide faster than nonexperts (e.g., Kundel et al. 2007). Acknowledged experts in the field of interest should not only be involved in the selection and development of clinical outcome measures but also take part in the rater training sessions. In addition to the training mostly focused on the proper use of study documentation, rater training may increase the awareness for the necessity to make explicit reasoning and documentation with regard to the diagnostic process and conclusion.

A further class of clinical scales, the self-reports often are of primary interest, apart from tolerance ratings, for the assessment of the efficacy of the treatment. Self-reports are the only method to measure and qualify symptoms, thoughts, beliefs, and opinions. Performance tests adopted to clinical conditions are important to complete the clinical outcome assessment. The use of performance tests measuring learning and

memory functions, attention, and problem-solving, for example, is in no way restricted to neuropsychopharmacology. These tests have shown to contribute to the understanding of CNS effects in various other indications (e.g., hepatic encephalopathy).

In one of their draft guidelines (2006) on patient-reported outcome (PRO) measures, the FDA stated that these measurement tools are of increasing importance in drug development. Self-reported questionnaires that are given directly to patients without the intervention of clinicians are often preferable to the clinician-administered interview and rating of the given answer. Questionnaires which are self-completed capture directly the patient's perceived response to treatment, without a third party's interpretation, and may be more reliable than observer-reported measures because they are not affected by interobserver variability (FDA 2006).

There are various topics which may be directly captured with the patient's response to a PRO. The choice of the item content, design, and response format will very much depend on the targets of the clinical program. A concept of PRO measures may be, for example, one of the following:

Discrete symptoms or signs, for example, frequency and/or intensity of pain sensation, and frequency of seizures

Overall conditions, e.g., depressiveness, fatigue, and general physical complaints

Feelings about the health condition, e.g., worry about disease getting worse

Feelings and opinions about the treatment, e.g., feeling the treatment is effective, and feeling relief of depressive mood since the start of treatment

General assessments, for example, improvement in physical functioning, treatment satisfaction, overall quality of life ratings, and health-related quality of life (HRQL) ratings

Specific assessments, e.g., decreased pain intensity and how bothersome specific symptoms are

Assessments of change after treatment (e.g., "symptom improved very much since the start of treatment"; "symptom improved most of the time after start of treatment," etc.) in contrast to absolute assessments

Clinicians have recognized that understanding the patient's perspective on the impact of disease and treatment on functioning and well-being is important for pharmaceutical, biologic, and medical device product development and evaluation. Clinical studies are increasingly incorporating health-related quality of life (HRQL) and other concepts of patient-reported outcome instruments into clinical trial programs for new drugs with the expectation that these outcomes will help inform physicians and patients on the beneficial effects of these treatments (e.g., Wilke et al. 2004). PROs may be further useful in differentiating the patient benefits among competing products with similar clinical efficacy and translating clinical effects into outcomes more meaningful to patients, relatives, and treating physicians. For a first classification of the various clinical scales, see Table 1.

Technically speaking, a diagnostic scale may be constructed as a questionnaire or a checklist asking for a step-by-step response from one item to the next or it may be used as a guidance for the experienced physician to conduct a patient interview (e.g., Hamilton Depression Scale). Diagnostic scales should reflect relevant diagnostic criteria in ICD-10 (http://www.who.int/classifications/icd/en/), other diagnostic classification systems, and/or medical or regulatory classification systems, e.g., DSM-5 for mental disorders. If available the scales selected for the program should also correlate with criteria from validated structured or semi-structured interviews established in the relevant medical speciality. The ongoing discussions about the scientific usefulness of diagnostic systems such as DSM or ICD-10 (e.g., Kendell and Jablensky 2003) may be taken into account. Both systems may not provide distinct classification in some disciplines, for instance, in psychiatry, that cannot always be transcribed into neurobiological pathways and genetic entities (cf. NIH Research Domain Criteria).

As already emphasized physicians, experts, and other professional personal responsible for

Table 1 A Taxonomy of the role of observer- and patient-rated scales in clinical development of drugs

Scope	Source of information	Content	Processing
Diagnosis; selection; exclusion	Physician/expert	Overall health status Symptoms, signs, or syndrome Functional status	Classification Identification Ranking
	Patient	Overall health status Symptoms/signs, individually or as a syndrome associated with a medical condition Functional status Activities of daily living Perception/worries about health Health-related quality of life	
	Relative/ caregiver	Functional status in daily situations Efforts in caring for the patient/carer burden	
Measurement of treatment outcome	Physician/expert	Clinical global impression of health status, severity of illness, change in condition, therapeutic effect, side effects, etc. (e.g., AMDP&CIPS 1990) Overall health status Health-related quality of life status and/or change Symptoms, signs, or syndrome Functional status	Effect, efficacy Benefit Tolerability, adverse events, estimation of Risk factors Covariate to primary effect measure
	Patient	Absolute and/or relative measure of change: Overall health status Symptoms, signs, syndrome Functional status Activities of daily living Perception/worries about health Patient satisfaction with the treatment and its results (e.g., Asadi-Lari et al. 2004) Health-related quality of life	
	Relative/ caregiver	Functional status in daily situations Carer burden	
Monitoring	Professional personal/ caregiver/relative	Documentation of study outside events	Nuisance variables, system description

assessments should be individually trained in the use of the diagnostic and observer-reported scales. Rater training should be an integrated part of any clinical explorative and pivotal trial to further valid diagnosis and reliable selection of study populations and generate informative co-variates for posteriori statistical adjustments and explorations in accordance to the study protocol.

Procedure

How to select, develop, and/or modify clinical scales eligible for the planned clinical program or a clinical study? Which rater (expert rater, observer, patient) will be informative on which level and in which domain of treatment effects? How to operationalize the research concept and to make hypotheses testable and find informative answers from explorative tests and questions?

The search for suitable clinical scales can be started with scanning guidelines and consensus publication from the medical speciality of interest. There are also critical reviews on scales available (e.g., AMDP&CIPS 1990). Usually, the first step to find appropriate scales will be to collect and compile items from various published scales (tasks, questions, or expressions); compare those items with regard to content, wordings, and response format; check for doubles; and decide on a first selection of items to be tested as draft item collection with experts in the field and ideally

with people who are familiar with symptoms and signs. The following criteria may be important for the selected items and scales check:

Wording
Comprehensibility
Completeness of topics
Response format
Availability of statistical item/scale characteristics.

These criteria are as well applicable to qualify standardized published scales. In case of foreseeable multicenter-multicultural studies, the planning team should consider translation of the eligible items/scales. This may, and mostly will, make linguistic validation and cultural transformations of concepts and cognitive debriefing necessary (Wild et al. 2009). These procedures are often very time consuming and should be planned right in time.

After compilation and adoption of the items according to the constructs, dimensions, and domains pre-specified in the clinical development plan, the scale(s) will be applied in a test sample of subjects (best choice would be a subsample from the target population of the developmental drug) in order to analyze item and scale characteristics and confirm or develop a scaling and scoring procedure ready for purpose. As discussed before, scaling is the part of measurement that involves the construction of an instrument that associates qualitative constructs with quantitative metric units. Scaling evolved out of efforts mostly in psychologic and educational research to measure constructs like intelligence, motivation, authoritarianism, self-esteem, and so on. In many ways, scaling remains a mostly misunderstood aspect of social research measurement. Remarkably it attempts to do one of the most difficult of research tasks, to measure abstract concepts, which have no obvious empirical conditions or which are not directly observable.

The *item* is regarded the unit of any scale (questionnaire, behavioral test). It is usually composed of three parts: (1) the test instruction (information on what the item or the complete scale is used for and how the choice for best response should be performed); (2) the presentation of a question, task, or an expression to which the respondent is invited to respond to; and (3) the choice of best response which is expected in the third part of the item offering an arrangement of words, numbers, idioms, symbols, or any other signal (response option of a specific format) to make best discriminative choice likely. The response format may be binary, multi-categorical, ordered categorical, or a response line with one or more anchors (visual analog scale, VAS). In some behavioral tests, the response is expected in a blank field (e.g., a gap at a certain position in a sentence where the respondent has to complete the sentence). Free form tests and questionnaires may collect interesting information. The free answers have to be separately analyzed and categorized to enable further statistical processing. In closed formats, the stimulus for choice of the best response should be well designed to induce a discrimination process reflecting the degree of agreement or disagreement.

The rating of the expression or question may in relation to the question be

Direct (e.g., are you satisfied with the results of this treatment: yes/no or another format with more than two categories)
Comparative (e.g., my headache is much better than before treatment: true/not true)
Magnitude (any frequency or intensity rating, e.g., for pain)

According to Stevens (1951), measurements in science are generally on four levels (see Table 2). The taxonomy may be extended by a fifth measurement level, the absolute scale (e.g., number of objects, probabilities).

Any numeric ordering or intuitive arrangement of verbal or symbolic response options does per se not reflect the true level of the measurement. In fact it has to be empirically shown that ratings, for example, on the frequency dimension with a typical arrangement like 0 = "never," 1 = "occasionally," 2 = "frequently," and 3 = "always" fulfil the criteria of ordinal or even interval level of measurement. Especially questionnaires in behavioral

Table 2 Levels of measurement and statistical processing

Level	Statistical description	Example	Test statistics
Nominal (or categorical)	Modus, frequency	Gender	Nonparametric tests
Ordinal	Median, percentile	Degree of agreement	
Interval	Arithmetic mean, standard deviation	Body temperature in Celsius or Fahrenheit	
Ratio	Geometric mean	Age, body weight	Parametric tests

and social sciences rarely reach the level of interval scale. Psychometricians are interested in scaling on a high level of measurement. Despite these considerations, the choice of the response format should always be guided by the content and scalability of the construct. Therefore, just attaching ordered natural numbers or percent numbers to design a "metric" response format is not scaling in a mathematical sense. The qualification of level of measurement of a scale has to be proven on response distribution characteristics based on empirical data.

Scales may further be divided into two broad categories: unidimensional and multidimensional. The unidimensional scaling methods were developed in the first half of the twentieth century, and some of them have been named after their inventor. Among the various scaling methods, the *psychophysical scaling* has a separate theoretical background. Psychophysics is a psychological discipline that has its roots back in the work of G. T. Fechner, E. H. Weber, and Wilhelm Wundt, founder of the first laboratory for experimental psychological at the University of Leipzig, 1879. Psychophysics deals with the relationship between physical stimuli and subjective correlates, in general the percept. Psychophysicists employ experimental stimuli that can be objectively measured, such as pure tones varying in intensity or lights varying in luminance or frequency. All the traditional senses have been studied including the enteric perception and the sense of time. Regardless of the sensory domain, three main procedures of investigation have been used: the definition of absolute threshold, discrimination threshold, and various scaling procedures using constant or systematically varied stimuli characteristics. The absolute threshold is the level of intensity or frequency at which the subject can just detect the presence of the signal. The difference threshold is defined as the magnitude of difference between two stimuli of differing intensity or frequency that the subject is able to detect. The just noticeable difference, also named difference limen (DL), is the difference in stimuli properties the subject notices with a defined proportion of the cases (mostly $p = 0.50$).

The determination of critical flicker fusion frequency and that of critical fusion frequency are examples of psychophysical measurement that have often been used and are still in use in psychopharmacology.

The *visual analog scale* (VAS) is in most cases a 100 mm horizontal line with two named poles or verbal anchors, like "not at all" and "very much" or similar wordings. The typical use of VAS is, for example, the following:

How severe was your pain today? Please place a mark on the line below to indicate how severe your pain was!

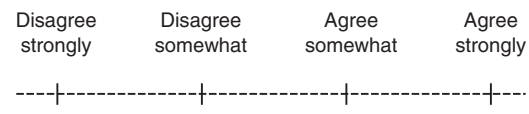
When using VAS as a measurement instrument, one tries to quantify a sensation, a trait, or any other entity on a ratio scale level assuming further that the entity's characteristics are ranging across a continuum from "none" to "very severe" or "very intensive" or within a similar concept. The assessment is highly subjective, in a practical way "imprecise," in regard to the positioning of the tic mark. VAS may be of value when looking at the intraindividual change of the entity, but they are most likely of less value for comparing groups. It has been argued that a VAS is trying to deliver interval or even ratio measures. But

there are no convincing arguments for the values being more than ordinal data. Despite the simplicity and "face validity" of this measurement method, data handling and interpretation have to be done with caution.

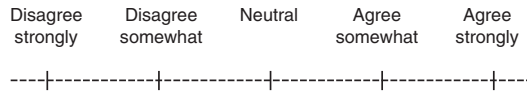
Thurstone scaling. Thurstone was one of the first scaling theorists. He invented three different methods for developing a unidimensional scale (e. g., 1927): the method of equal-appearing intervals, the method of successive intervals, and the method of paired comparisons. The three methods differed in how the scale values for items are constructed, but in all three cases, the resulting scale is rated the same way by respondents. The method of equal-appearing intervals is explained as it is the easiest method. Because this is a unidimensional scaling method, the concept one is trying to scale is reasonably thought of as one-dimensional. When starting the procedure, the description of this concept should be as clear as possible so that the persons who are going to create the statements (items) have a clear idea of what the investigator is trying to measure. Next, the developer will ask people to generate similarly worded statements about the concept. Then the participants are asked to rate each statement on an 11-point response scale with a predefined criterion, like how favorable the statement appears to them with regard to the construct. Next the ratings will be analyzed. For each statement, one needs to compute the median and the interquartile range. The median is the value above and below which 50% of the ratings fall. The first quartile (Q1) is the value below which 25% of the cases. The median is the 50th percentile. The third quartile, Q3, is the 75th percentile. The interquartile range is the difference between third and first quartile, or Q3–Q1. To facilitate the final selection of items for the scale, one might write the parameters into a table; maybe we want to sort the statements in the table of medians and interquartile range in ascending order by median and, within that, in descending order by interquartile range.

For the final scale, one should select statements that are equally distributed across the range of medians. Within each median value, try to select the statement that has the smallest interquartile range. Once the items for the final scale are selected, one should test the new scale in an independent sample of eligible patients or members of the target population.

Likert scaling is a unidimensional scaling similar to Thurstone scaling. The term "Likert scale" is used in various ways in literature. Sometimes the term actually seems to describe Likert or Likert-like items. Likert scales are the four- to nine-point scales much used in clinical trials and in many other fields of research. The scale is often used in a semantic polar format, that is, in relation to the statement or question, the response options correspond to wordings like "agree strongly" and "disagree strongly." Example:



The five-point scale is probably the most commonly used response format version, which in many cases in contrast to the four-point scale will include a midpoint "neutral".



The assumption with the VAS as well as behind numeric rating scales, including Likert scale, is that the geometrical distance between markers on a line or of tic boxes in combination with verbal expressions and/or numbers is homologous. This is one of the reasons why the graphical layout of those items should guarantee equally spaced elements of the response format.

After definition of the concept, verbalization of the statements, tests of the draft version, and confirming those items that form a reasonable scale, the final score for the respondent on the multi-item scale is the (weighted) sum of their ratings on all items (sometimes called a "summated" scale). On some scales, one will have items that are reversed in meaning from the overall direction of the scale. To cumulate item scores, one will have to inverse the response score of this item to maintain unidirectional scaling. Likert scales, like other item scales, may be problematic in comparison across groups. The expectation of a

researcher would normally be that the mean response will vary across treatment groups. The problem is that in many cases the variances will also differ. The variance has to be less at the ends of the scale, as there is no alternative response to one side of the endpoint. For example, with a five-point scale, the variance would be expected to be largest at the midpoint, 3, and smallest at the extremes. A possible solution to this problem might be to use the arc sine square root transformation of the scores. The responses are divided by 5, to yield a number between 0 and 1. The square root is taken (still between 0 and 1). The angle whose trigonometric *sine* is that number is the transformed response and can be used for further statistical analysis.

Osgood et al. (1957)'s *semantic differential* was designed to measure the connotative meaning of concepts. The respondent is asked to choose his or her position on a scale between two bipolar adjectives (e.g., "adequate–inadequate" or "valuable–worthless").

Sometimes it may be difficult to find properly defined poles of the differential! Therefore, many researchers prefer unipolar item scales (e.g., mood scales and multi-item pain scales).

Guttmann scaling. This method is also known as cumulative scaling. Like with the other examples of item response scaling, this method starts with the definition of the construct of interest, the generation of a large set of statements that are judged by some experts or members of the target group how favorable (yes/no rating) the expressions are in regard to the construct. Following this, one constructs a matrix or table that shows the responses of all the respondents on each of the items. Afterward this matrix is sorted so that respondents who agree with more statements are listed at the top and those agreeing with fewer are at the bottom of the matrix. For respondents with the same number of agreements, the statements are sorted from left to right from those that most agreed to those that fewest agreed to. In case of only a few items, one can easily examine this matrix. In larger item sets, the method of choice may be the *scalogram analysis* to determine the subsets of items from our pool that best approximate the cumulative property. After the review of these items follows the selection of the final scale elements. In many cases, there is no perfect cumulative scale and the researcher will have to test for goodness of fit. These statistics will estimate a scale score for each of the items that are used in the calculation of a respondent's score.

In the late 1950s and early 1960s, measurement theorists developed more advanced techniques for creating multidimensional scales. *Multidimensional scaling* (*MDS*) is a data reduction technology normally using a direct similarity or dissimilarity matrix. MDS fits a set of points in a space such that the distances between the points are as closely as possible to a given set of dissimilarities between a set of objects, for example, ratings. MDS does not make distribution assumptions necessary. As MDS is a spatial method, there are metric assumptions, for example, the distance from some point A to B shall be the same as from B to A. This might sound strange to the reader, but in some situations, two points A and B may not be bidirectionally equidistant. Consider, for instance, the distance between home and work, which may be due to specific situations in the morning and the evening not of identical length. If the equidistance assumption cannot be fulfilled, one should not use MDS. Anyway, the decision whether to use or construct one- or multidimensional scales depends very much on the clinical target (will efficacy in one dimension clinically be sufficient or will relevant drug effect be expected in more than one dimension) as defined in the clinical development plan. If the construct is one dimensional, one also will use one-dimensional scales, and if the construct is of known multidimensionality, one should consider multidimensional scales or several one-dimensional scales. Both ways will offer their special advantages and disadvantages with regard to the upcoming point of decision on the clinical trial outcome.

The application of mathematical models to response data from questionnaires, clinical, educational, and psychological tests, is discussed and described in test theory. Test theory is a body of theory that offers mathematical models to make statistical adjustments in response data in order to predict, describe, or estimate a person's trait, ability, attitude, or any other construct. There are in

general two different test theories, which are of relevance in the present context and which help to understand the steps from single item or item pool generation to item construction, definition of adequate response options to testing the first draft questionnaire, and confirmation of the final measurement instrument. The methods are also very helpful for the reevaluation of known scales and items. The two main test theories are:

1. *Classical test theory*, which assumes that for each person, we have a true score of some ability or characteristic, T, which would be obtained if there were no errors in our measurement. Because instruments used for measurement (and sometimes the users of those too) are imperfect, the score that is observed for each entity, for example, a person's ability, most times is different from the person's true abilities or attitudes. It is concluded that the difference between the true score and the observed score is the result of measurement errors. Classical test theory is dealing with the relation of the true score T, the error E, and the observed score X. Formally:

$$X = T + E.$$

Further assumption: True score T and error E are not correlated, $\rho(T,E) = 0$.

The most important concept is that of reliability. The reliability of the observed test scores X, denoted as $\rho^2(X, T)$, is defined as the ratio of true score variance $\sigma^2 T$ to the observed score variance $\sigma^2 X$. Because it can be shown the variance of the observed scores to equal the sum of the variance of true scores and the variance of error scores, it follows that

$$\rho^2(X,T) = \frac{\sigma^2 T}{\sigma^2 X} = \frac{\sigma^2 T}{\sigma^2 T + \sigma^2 E}.$$

The reliability of test scores becomes higher as the proportion of error variance in the test scores becomes lower and vice versa. The reliability is equal to the proportion of the variance in the test scores that could be explained if we knew the true scores. The square root of the reliability is the correlation between true and observed scores.

2. *Item response theory* (*IRT*), also known as latent trait theory, is a set of probabilistic models and the application of mathematical models to response data from behavioral tests and questionnaires measuring abilities, characteristics, or other variables. IRT models apply functions to quantify the probability of a discrete outcome, such as a correct response to an item, in terms of *person and item parameters* (see, e.g., Linden and Hambleton 1997; Rost 2004).

   Person parameters may, for example, represent the cognitive ability of a patient or the severity of a patient's symptom. Item parameters may include item difficulty (location), discrimination (slope), and random guessing (lower asymptote). IRT does not only apply to discrete binary data but may also deal with ordered categorical data to indicate level of agreement and other response dimensions as already discussed earlier. One of the purposes of IRT is to provide a framework for evaluating how well assessments and individual questions on assessments work. In drug development programs, IRT may be very helpful to collect and construct items and maintain item pools for clinical trials in a defined indication and develop or adopt new scales within the conceptual framework of the clinical program.

The performance of an item in a test is described by the *item characteristic curve* (ICC). The curve gives the probability that a person with a given ability level will answer the item correctly or give an answer in line with the expectations according to the construct definition. Persons with lower ability ($\theta < 0.0$) have less of a chance to answer correctly or agree on a yes/no item, while persons with high ability are very likely to answer correctly.

IRT models can be divided into two families: one-dimensional and multidimensional models. One-dimensional models require a single trait (e.g., ability) dimension θ. Multidimensional IRT models analyze response data arising from

multiple traits. However, because of the greatly increased complexity with increasing number of included traits, the majority of IRT research and applications utilize a one-dimensional model. The models are further on named according to the number of parameters estimated. The one-parameter logistic model (1PL) assumes that there is only minimal guessing by the respondent and that items have equivalent discriminations, so that items can be described by a single parameter (b_i). The 1PL uses only b_i, the 2PL uses b_i and the parameter a_i, and the 3PL uses b_i, a_i, and item parameter c_i.

A given model describes the probability of a correct response (or a yes/no response option where one is defined as correct and the other as incorrect in the frame of some syndrome or disease theory) to the item as a function of a *person parameter*, which is in the case of multidimensional item response theory, a vector of person parameters. For simplicity we will stay with the model of only one person parameter. The probability of a correct response depends on one or more item parameters for the item response function (IRF). For example, in the three-parameter logistic (3PL) model, the probability of a correct response to an item i is given by

$$p_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta - b_i)}}$$

where θ signifies the person parameter, e is the constant 2.718, and a_i, b_i, and c_i are the item parameters.

As can be seen from Fig. 1, the item parameters simply determine the shape of the IRF. The figure depicts an example of the 3PL model of the ICC with an explanation of the parameters. The parameter b_i represents the item location (item difficulty). It is over the point on θ where the IRF has its maximum slope. The simulated example item is of medium to higher difficulty, since $b_i = 1.3$, which is to the right of the center of the distribution. The item parameter $a\ i$ represents the (rather good) discrimination, the degree to which the item discriminates between persons in different regions on the latent continuum. This item parameter characterizes the slope of the IRF where the slope is at its maximum. The item parameter $c_i = 0.20$ indicating that persons with low ability may endorse correct response.

One of the major contributions of item response theory is the extension of the concept of reliability. Traditionally, reliability refers to the precision of measurement (i.e., the degree to which measurement is free of error). And traditionally, it is measured using a single index, such as the ratio of true and observed score variance (see above). This index is helpful in characterizing an average reliability. But IRT makes it clear that precision is not uniform across the entire range of test scores. Scores at the edges of the test score range generally have more error associated with them than scores closer to the middle of the range.

Item response theory elaborated the concept of item and test information to replace reliability. Information is also a function of the model parameters. According to Fisher information theory (named after the inventor and famous statistician R.A. Fisher), the item information supplied in the case of the Rasch model (Rasch 1960) for dichotomous response data is simply the probability of a correct response multiplied by the probability of an incorrect response:

$$I(\theta) = p_i(\theta)q_i(\theta).$$

The standard error (*SE*) is the reciprocal of the test information at a given trait level:

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

Intuitively, one can agree to the conclusion that more information implies less error of measurement.

After this short excursion into some basics of test theory, one may agree that measuring is not just assigning numbers to empirical objects or events (see the straightforward definition of S.S. Stevens 1951). In the classical definition, measurement is the estimation of ratios of quantities. Quantity and measurement are mutually defined: quantitative attributes are those which make measuring possible. In terms of representational

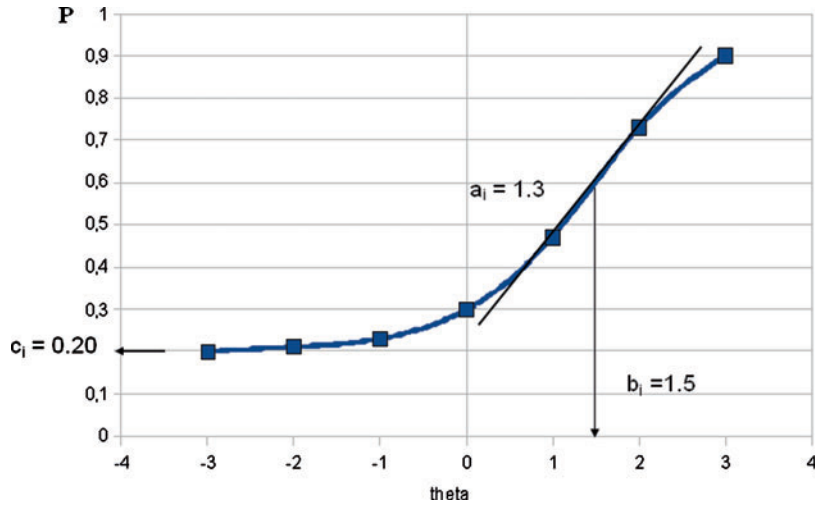


Fig. 1 IRF for a hypothetic item data with item parameters $b_i = 1.5$, $a_i = 1.3$, and $c_i = 0.2$

theory, numbers are assigned based on similarities between the structure of number systems and the structure of qualitative systems. A property is quantitative if such structural similarities can be established. This definition is much stronger than the definition of Stevens (1951).

Evaluation

Drug development in known or new indications will start with an extensive literature search for measurement models and scales in the target area. When searching and selecting suitable scales for the clinical development plan, essential information about the scale characteristics and properties are of major importance. Information with regard to the following topics may be needed:

Completeness and representativeness with regard to the concept of interest
Relation to medical and mathematical measurement models available
Published indices or data for at least scale reliability and validity
Sufficient evidence on satisfying scale properties
Evidence for validated linguistic and/or cultural versions, as in many cases pharmacological drug development will increasingly often be performed in multicultural and multilingual studies

In addition instructions for the standardized application of the test or questionnaire, procedures of training raters for the proper use of those instruments are as important as the mostly cited scale properties reliability and validity. Further information about the way the items shall be presented, the scoring rule, for which experimental conditions and in which population the indices and coefficients are valid is needed. It is almost always necessary to consider some kind of reevaluation of the selected scale(s) for the own new project. Clinical scales used as outcome assessment tools are key elements in patient-focused drug development. Their reasonable use is recommended not only by the authorities in the USA and Europe.

The necessity of a conceptual framework before starting the clinical development program of a drug is known for decades in the field of pharmacodynamic research and development. It has been explicitly named in the guideline for patient-reported outcome measures (FDA 2009). The conceptual framework should combine three major concepts:

1. Treatment or interventional concept
2. Target population for the intended treatment
3. Measurement concept (including the endpoint model)

Figure 2 gives an overview of some important steps in the developmental process from the

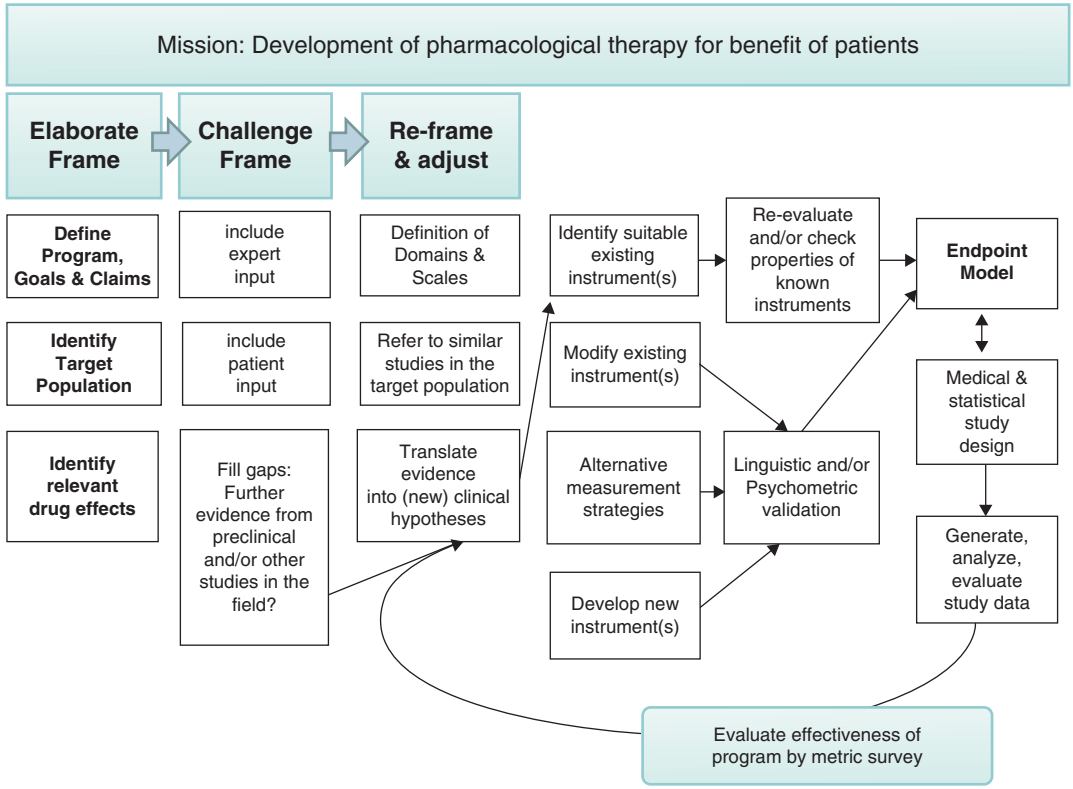


Fig. 2 Critical path description from concept development to endpoint model. The *arrows* shall indicate the flow of information during the process

measurement and scaling concept until the final definition of the endpoint model and the entrance of these findings and decisions into a clinical study protocol. These steps can be planned and performed in a way of inductive and deductive reasoning including strategic decision nodes. Suffice to say scales can be used to document and evaluate these processes in innovative drug development. Theories of judgment and decision making are interesting fields, but their discussion is beyond the scope of this chapter. The interested reader may find interesting topics in Kahneman (2011).

Critical Assessment of the Method

The selection of scales for a drug development program should always be guided by the program goals and claims that are intended to reach for. Sometimes the decision to select and use known scales seems to be influenced by their availability, their application in similar clinical programs or clinical studies, and published data. In case of diagnostic scales, this strategy may be appropriate as long as the drug development program is within a well-known traditional indication and patient population. On good reason to rely on established scales is the comparability of new results with former study outcomes. Innovative drug development, e.g., based on a new mode of action, or intended for special populations (e.g., children, or very old persons), will, in many cases, make modification and/or development of diagnostic scales, scales for inclusion criteria assessments, and above all outcome measures necessary. Some of the following criteria in case of search and selection of known scales may be helpful for the decision. For what purpose and according to which clinical and methodological framework

was the scale developed did scale development care for sufficient content validity? Does the item concept (question, task, or statement, and the response format) match the intended construct to be investigated? Is the scale a single-item or multi-item scale and does it appropriately fit the one-dimensional or a multidimensional construct? In the preceding paragraphs, some information about item and scale development methods have been discussed which might be of additional help to answer these questions during search, selection, evaluation, and decision on eligible scales for the planned clinical program.

Critical *item analysis and construction* will include semantic and mathematical methods as well as consideration on the design of item and/or scale. Further the planning team may consider more modern technical solution to present items and scales (e.g., tablets, smartphones, WEB-based surveillance platforms) which could offer new perspectives on drug assessment. These new technologies offer opportunities to "continuously" gather data (e.g., dairy data). Privacy and data protection, validation of the system, testing the validity of items and scales, and planning adequate data handling and analysis will be a challenge. The computerized presentation of clinical scales will need some attention on wording, layout, and arrangement of response options to direct the respondent's attention to the task or question and to encourage to an open unbiased response. There is a considerable number of literature which reports on research issues and findings of psychologists and survey methodologists on the cognitive, motivational, and communicative aspects of presentation and comprehension of items and scales (for overview, e.g., Schwarz 1999). Most items are composed of three parts, the test instruction (and presumably further information on the context of the surveillance), the question/task itself, and the response part which may be designed in open or closed format. Open formats would allow the respondent to give free answers, but in clinical research (except for questions on tolerability and adverse events), probably the most preferred approach, especially when asking for efficacy of treatment (as primary outcome measure), is the closed format. Items researchers will have to consider the specific cognitive, motivational, and communicative processes underlying task or question and the dimension and response format under clinical conditions. Unlike measurement in natural sciences, behavioral measurement is based on complex cognitive and emotional processing of the information given to the respondent. Patients, and also doctors, tend to draw information from the content and design of an item and may determine more or less explicitly what the best answer or solution would be. Experimental conditions and scales provide some information to the respondent which can result in unexpected outcomes. Therefore, the wording of the statements or expressions should always be comprehensible and readable to the intended respondent. In case of diagnostic scales, the items must be coherent to relevant criteria of the disease or the syndrome. Patient-reported scales should avoid as much as possible technical or medical terms and replace them by more daily speech terms. The input of patients from the target population is mandatory to find appropriate speech. In new fields structured interviews may be more informative than unstructured interviews. With regard to the response options and the response scaling, the response scale should reflect both temporal properties of the disease or symptom(s) as well as the temporal aspects of the hypothesized time and mode of action of the drug, all issues be adequately reflected in the study design and protocol. Typical response options used in clinical trials are on the intensity or frequency dimension with ratings related to time windows like "yesterday," "this morning," "the last week," "the last month," "since the last visit," and others. The choice of temporal anchoring of assessments must be in line with the nature of the symptom or the disease and must consider reasonable memorization of the construct, symptom, or process. Answers about the past are less reliable when they are very specific, because the questions may exceed the subjects' capacity to remember accurately. There is much literature about the functioning of autobiographic memory (e.g., Williams et al. 2008). Human judgment is context dependent by drawing on the information that is most accessible at the time point of

measurement. In certain cases the respondent may tend to randomly guess or feel encourage to respond in direction of a presumed desired direction. As a practical example, items asking for comparative judgment, for example, "How is your pain today," "no pain," "pain as usual," or "pain more severe than usual," are obsolete because of the undefined anchor "usual." When drafting items, the researcher must be focused on the wording (1) to match theoretical criteria or research object of interest and (2) be aware of item characteristics interpreted as relevant information by the respondent. Comprehension, interpretation, the type of memorizing the topic, and communicative information of the item (and the scale in total) are examples of cognitive processing of questions, statements, and tasks and deserve special attention.

Implementation of a scale into the clinical development plan and single studies. There is a common misconception that if a published validated instrument is chosen for a clinical study or evaluation program, one does not need to check the reliability and validity of the instrument in the target population. If content and response format do not fit into the study design, the planning team should consider modifications. The modified instrument has to be tested again for its scale properties in a separate sample taken from the target population before using it in explorative or pivotal clinical trials. If the investigated symptom has temporal properties in the way that its appearance or intensity is changing during daytime, or its appearance or intensity may vary from day to day, the investigator might consider the implementation of a patient's diary in the study design. FDA supports the use of diaries where appropriate, but "If a patient diary or some other form of unsupervised data entry is used, the FDA plans to review the protocol to determine what measures are taken to ensure that patients make entries according to the study design and not, for example, just before a clinic visit when their reports will be collected" (Federal Register Vol. 71, Nr. 23, pp. 10; 334–337). In recent years, systems for use of electronic questionnaires, PRO, and diaries frequently offered by specialized companies are used by sponsors to facilitate data selection. The eventual advantages of the administration of well-controlled electronic questionnaires may be among others (e.g., Kelly 2015)

- Help to counteract declining response rates
- Reduce collection costs
- Improved data quality
- Rapid and continuous access to data
- More detailed insight into subject's behavior regarding input of data

Validation, reliability, and compliance with Part 11 electronic data requirements (http://www.fda.gov/ora/compliance_ref/Part11/) are important issues that will be raised by the FDA and other authorities, and the sponsor will have to successfully address them. A typical question is often heard: Are paper questionnaires and electronic questionnaires equivalent? The answer most likely is no, because any modification in a scale (including method of administration) will certainly need some additional validity evidence. *Scale properties*. A questionnaire or test as an instrument to measure entities or a construct should have specific properties in order to be accepted as a scale. The instrument should be standardized in the way and the circumstances it shall be presented and used, the evaluation of the ratings shall be defined and objective in a way that the results of this evaluation is independent of the evaluating person, and there should be important data available on its reliability and validity. In psychology, the concept of validity has in addition to the already discussed aspects two further fields of application. The first aspect addresses test validity, a concept that has evolved in psychometrics, dealing with theory and technique of psychological and educational measurement. The second is related to the study design, pointing to the fact that different types of studies are subject to different types of bias. For example, recall bias is likely to occur in cross-sectional or case-control studies where subjects are asked to recall exposure to life events or other special events. Subjects with the relevant condition (e.g., the degree of disease or syndrome to be investigated) may be more likely to recall relevant events that they had experienced

than subjects who do not have the condition or have a lower degree of it.

In contrast to test validity, assessment of the validity of a research design does not involve data collection or statistical analysis but rather evaluation of the design in relation to the desired conclusion on the basis of prevailing standards and theory of research design. This obviously is an issue in drug development programs, especially when designing several phase III studies in various regions and cultures.

Test validity, which is in focus here, can be assessed in a number of ways (APA 2014). Test validation typically involves more than one type of evidence in support of the validity of a measurement method (e.g., structured interview, questionnaire, test, etc.). The various types of validity include content-related, construct-related, and criterion-related evidence with the subtypes concurrent and predictive validity according to the timing of the data collection. In the following we will present and discuss some of the various aspects.

Construct validity evidence involves the empirical and theoretical support for the interpretation of the construct. A good construct validity has a theoretical basis which is translated through clear operational definitions involving measurable indicators. Construct validity evidence includes statistical analyses of the internal structure of the test including the relationships between responses to different test items. They also include relationships between the test and measures of other constructs. Researchers should establish both of the two main types of construct validity, *convergent* and *discriminant*, for their constructs.

Convergent validity is assessed by the correlation among items that make up the scale or instrument measuring a construct (internal consistency validity); by the correlation of the given scale with measures of the same construct using scales and instruments proposed by other researchers, if appropriate, with already accepted in the field (criterion validity); and by correlation of relationships involving the given scale across samples.

Internal consistency is one type of convergent validity that seeks to assure there is at least moderate correlation among the indicators for a concept. Cronbach Coefficient Alpha is commonly used to establish internal consistency (as well as an aspect of reliability and for evidence of construct validity) with at least Alpha of 0.60 considered acceptable for exploratory purposes, Alpha of 0.70 considered adequate for confirmatory purposes, and Alpha of 0.80 considered good for confirmatory purposes.

Simple factor structure is another test of internal consistency, seeking to demonstrate for a valid scale that indicator items for a given construct load unambiguously on their own factor. This tests both convergent and discriminant validity.

Rasch models, one-parameter logistic models (1PL), are also internal consistency tests used in item response theory for binary items. Rasch models for polytomous items are also available. They are generalizations of 1PL Rasch model. Like Guttman scales, Rasch models test that the included items which are measuring a construct will form an ordered relationship (see Rasch 1960). A set of items may have ordered internal consistency even though they do not highly correlate (additive internal consistency as tested by Cronbach Alpha or factor structure). Ordered internal consistency reflects the difficulty factor, which means that correct response to a more difficult item will predict the response on less difficult items but not vice versa.

When factor analysis is used to validate the inclusion of a set of indicator variables in the scale for a construct, the researcher is assuming a linear, additive model. Linearity is assumed as part of correlation, which is the basis for clustering indicator variables into factors. With additivity is meant that items will be judged to be internally consistent if they are mutually highly correlated. However, items may lack high intercorrelation but have a strong ordered relationship. For this reason, many researchers prefer to use a Rasch model for scale construction, in preference to additive models like Cronbach Alpha or factor analysis.

Discriminant validity, the second major type of construct validity, refers to the principle that the indicators for different constructs should not be highly correlated. Discriminant validity analysis refers to testing statistically whether two

constructs differ as opposed to testing convergent validity by measuring the internal consistency within one construct. In constructing scales, some researchers reject an indicator if it correlates more highly with a construct different from the one which was intended to be measured. Some researchers use $r = 0.85$ as a rule-of-thumb cutoff value for this assessment. Construct validity is not distinct from the support for the substantive theory of the construct that the test is designed to measure, which is an issue for measurement models in drug development. Experiments designed to reveal aspects of the causal role of the construct may contribute to construct validity.

Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. Content-related evidence typically involves subject matter experts evaluating test items against the test specifications. Content validity is also called *face validity* and has to do with items seeming to measure what they claim to do. In content validity, one is also concerned with whether the items measure the full domain implied by their label. Failure of the researcher to establish credible content validity may easily lead to rejection of his or her findings. For help one should consider the use of surveys of panels or experts and/or additional focus groups of representative subjects. In case of PROs, asking patients, are ways in which content validity may be established.

It is a challenging task to make sure that the measures operationalized by experts or common sense sufficiently address the concept of the later scale. There could also be a *naming fallacy*. Indicator items may display construct validity, yet the label attached to the concept may be inappropriate.

Criterion validity evidence involves the correlation between the test and a criterion variable (or several variables) taken as representative of the construct. The correlation with known and accepted standard measures or criteria is of interest. Ideally these criteria are direct objective measures of what is being measured. Where direct objective measures are unavailable, the criteria may be merely closely associated. For example, employee selection tests are often validated against measures of job performance. If the test data and criterion data are collected at the same time, this is referred to as *concurrent validity* evidence. If the test data is collected first in order to predict criterion data which is collected at a later point in time, then this is referred to as *predictive validity.*

Reliability. According to classical test theory, reliability is not a fixed property of a test but a property of test scores that is relative to a particular population. A reliability coefficient is computed for a sample. This is because test scores will not be equally reliable in every population or even every sample. For instance, as is the case for any correlation, the reliability of test scores will be lowered by restriction of range. Also note that test scores are perfectly unreliable for any given individual *i*, because, as has been noted above, the true score is a constant at the level of the individual, which implies it has zero variance, so that the ratio of true score variance to observed score variance, and hence reliability, is zero. The reason for this is that, in the classical test theory model, all observed variability in *i*'s scores is random error by definition (see above). Classical test theory is relevant only at the level of populations and samples, not at the level of individuals. Reliability cannot be estimated directly since that would require one to know the true scores, which according to classical test theory is impossible. Estimates of reliability can be obtained by various means. However, there is no one standard method. The method of assessing reliability must reflect the medical use of the instruments. Some of the statistical and psychometrical methods are as follows.

Frequently the *p*-value is cited as evidence of reliability: a significant Pearson correlation means a correlation significantly different from 0. But one should scatterplot the data and check for biased values. The concordance correlation coefficient addresses the concept of agreement. However, it can be misleading in that it summarizes the fit around the line of identity and, therefore, like the Pearson correlation, a value close to one may not denote lack of variability around the line.

If a cut point is to be used to classify patients, agreement of the classifications could be

examined, using Kappa indices. Kappa is commonly used to measure reliability or agreement for nominal or ordinal variables; however, it also has limitations. If one method is a gold standard, then predictivity (sensitivity, specificity, or similar statistics) should be determined. Receiver operating characteristic (ROC)-type analyses have much to offer; also it can be argued that paying attention to the misclassifications, rather than the consequences of misclassification, may not result an appropriate comparison (Obuchowski 2005).

The intraclass correlation coefficient or its analogs (Bland and Altman 1996) is another class of models to test the consistency of ratings made by different observers when rating the same entity. However, the value of this method depends heavily on the sample used, and without repeated measurements, estimates of precision are impossible.

Stability of the response. The same form of a test is given on two or more separate occasions to the same group of examinees (test–retest). On many occasions, this approach is not practical because repeated measurements are likely to make changes within the rater (patient or observer). For example, the rater could adapt the test format and thus tend to score higher in later tests. A careful implementation of the test–retest approach is recommended. If appropriate and possible, parallel-test forms of the scale will be of great help in case of repeated measurement, which is the rule in most clinical trials, to control, for instance, memory and/or training bias. Extensive training of observers before entering the clinical trial is another method of reducing this kind of potential bias.

An aspect of reliability of special interest in drug development is the instrument's *sensitivity to change (responsiveness)*. In more general terms, it is that the measured scores are changing in direct correspondence to actual changes in the entity under treatment. There is a growing recognition that assessing the effect of an intervention should not only focus at the statistical significance of the differences in outcome measures between the experimental and the control group but should also focus at the relevance or importance of these outcomes. Estimating the magnitude of the difference between change scores in both groups, the difference between mean change scores may be expressed in standard deviation units with the effect size index (ES). One of the possible definitions has been developed by Cohen. Unfortunately, there is no agreed standard method for the estimation and the interpretation of the magnitude of intervention-related change over time or responsiveness assessed with outcome measures. For further details, see Middel and van Sonderen (2002) who are discussing advantage and limitations of several ES proposals.

Form equivalence is related to two or more different forms of test or questionnaire (sometimes called parallel version) based on the same content and administered in an identical way to the respondent. The presentation of a test (or questionnaire) one time as a paper–pencil test version and the next time as a computer-based test version is not regarded as being parallel versions and cannot be exchanged in a setting assuming equally valid and reliable. After alternate/parallel forms have been developed, they can be used for different persons or for several measurement occasions with the same person in a trial. This method is, for instance, very common in educational examinations to prevent communication between participating people. A person who took form A earlier could not share the test items with another person who might take form B later, because the two forms have different items. We should always consider the use of parallel-test versions in trials with intraindividual repeated measurements when we cannot exclude considerable training effect or change in the strategy of responding to the items caused by experience with the test.

Internal consistency is defined as the association of responses to a set of questions designed to measure the same concept. It is normally expressed by the coefficient of test scores obtained from a single test or survey. Usually, internal consistency is measured with Cronbach Coefficient Alpha, or its algebraically equivalent, the Kuder-Richardson Formula 20, when the data are dichotomous, or the Spilt-half method based on the assumption that two halves of a test is parallel except for having different variances.

Cronbach Alpha, which is the most frequently used and easily available procedure in nearly every commercial, statistical software package, is defined by:

$$\alpha = \frac{n\overline{r}}{(1 + \overline{r}(n-1))}.$$

Here the Coefficient Alpha is based on the average size of item-to-total score correlations, sometimes named standardized Alpha. One could also use the item-to-total score covariances that may be more informative when the items have different variances.

To describe the logic of internal consistency more vivid, assume patients participating in a postmarketing survey about drug D indicated for treatment of symptom S. They are asked to rate statements about their satisfaction with the treatment. One statement is "Drug D helped me very much in getting rid of the symptom." A second statement is "After intake of the drug I frequently experienced unusual headache." A third statement is "If the symptom will come back, I will use drug D again." People who strongly agree with the first statement would most probably agree with the third statement and vice versa. Patients, who agree with the second statement, will most probably disagree with statement one and, depending on the anticipated need for future treatment, or the availability of alternative therapy, will more or less disagree with the third statement. If the rating of the statements is patternless high and low among the participants of the survey, the responses are said to be inconsistent. When no pattern can be found in the patients' responses, probably the test is too "difficult" and patients just guess the answers randomly. Of course, different conclusions could be drawn from inconsistent results, like, the items may be reworded or items addressing similar aspects of patient's satisfaction may be added to the survey to capture the intended construct in a more reliable way. Internal consistency is a measure based on the correlations (or covariances) between different items or statements on the same questionnaire or test. It measures whether several items that presumably measuring the same construct are producing similar scores. The procedure of Cronbach Alpha is a statistic calculated from the pairwise correlations between items. The coefficient ranges between 0 and 1. In case where some or many items are negatively correlated with the total score, the coefficient can take on negative values even less than −1.0. One can check the effect of those items by reversing the item scoring and run the procedure again. As a rule of thumb, Alpha of 0.6–0.7 indicates acceptable reliability and Alpha of 0.8 or higher indicates good reliability. High reliabilities (0.95 or higher) are not necessarily desirable, as this indicates that the items may be entirely redundant. The goal in designing a reliable instrument is for scores on similar items to be related (internally consistent), but for each to contribute to some part a unique information.

In 2004, Lee Cronbach, the inventor of Coefficient Alpha as a way of measuring reliability, reviewed the historical development of Alpha: "I no longer regard the formula as the most appropriate way to examine most data. Over the years, my associates and I developed the complex generaliability (G) theory" (Cronbach 2004, p. 403). Discussion of the G theory is beyond the scope of this contribution. Cronbach did not object the use of Coefficient Alpha, but he recommended that researchers should take the following into consideration while employing this approach:

- Standard error of measurement is the most important piece of information to report regarding the instrument, not a coefficient.
- Independence of sampling.
- Heterogeneity of content.
- How the measurement will be used: Decide whether future uses of the instrument are likely to be exclusively for absolute decisions, for differential decisions, or both.
- Number of conditions for the test.

Ratings repeatedly performed on the same entity or object by one observer ask for reliability of the rater's judgment, which is called intra-observer or *intra-rater reliability*. The comparison

between the rating of several raters on the identical entities (objects, persons, etc.) is called the interobserver or *inter-rater reliability*. Statistical methods for measuring agreement between categorical outcomes are well established. Cohen (1960) developed the kappa statistic as an agreement index for two binary variables. It has an appealing interpretation as a measure of chance-corrected agreement. Later, Cohen (1968) generalized the original kappa to the weighted kappa coefficient for ordinal discrete outcomes. Since its development, kappa with its extensions (Cohen 1960, 1968; Fleiss 1971, 1981; Fleiss and Cohen 1973 and others) has been well studied in the literature and broadly applied in many areas.

From the previous presentation and discussion, one may draw the conclusion that after considerable discussion about the scientific value of validity evidence and the relation between reliability and validity, the message is that reliability is a *necessary* but not *sufficient* condition for validity.

Modification of the Method

The psychometrical methods of development, evaluation, and application of measurement tools may be even more useful for drug development if they are integrated part in the conceptualization of treatment, target population, and measurement in an early phase of drug development. Scales may then be more specifically selected, modified, or developed to depict essential aspects of person, disease, and interventional properties and are no longer restricted to the role of patient selection criteria or endpoint definitions for single clinical studies.

The commitment to plan drug development programs instead of isolated experiments and/or trials and start modeling of interventional effects right from the beginning of the program lets one to also think of big data analytics. Big data analytics (BDA) is defined as the application of advanced (exploratory) analytic techniques to very big data. The term big data is used for data of high volume with diverse data types and/or velocity (streaming data) (Gartner IT Glossary; http://www.webopedia.com/TERM/B/big_data_analytics.html). The analytical techniques comprise, e.g., predictive analytics, data mining, artificial intelligence algorithms, and language processing. The potential of an IT approach to drug development programs make installation of platforms suited to BDA necessary.

Advanced IT has influenced the way scales are presented and processed in experiments and clinical trials. As said before, "scale" should no longer be defined in the traditional way as a question or task that is physically presented as paper–pencil–tool or established on a wired apparatus with a monitor and keyboard. Scales are increasingly frequently presented to the responders as computerized survey on smartphones, tablets, or WEB-based on computers. The computerized versions of scales need special investigation in the validity, reliability, and eligibility of the tool and demand new concepts of statistical data analysis. Integrated measurement models are in line with new concepts programs of modern drug development (e.g., EMEA/127318/2007).

The view and opinion of patients (and relatives or health carers) on treatment outcomes will increasingly contribute to the development of new, effective, and safe medicinal drugs and translate pharmacological and clinical study outcomes into meaningful information for physicians, patients, and their relatives in everyday practice. Measurement concepts extended beyond the clinical disease models, and related target populations may then enable links from concepts of preclinical experiments to cost–benefit quantifications of marketed drugs.

References and Further Reading

Acquadro C, Jambon B, Ellis D, Marquis P (1996) Language and translation issues. In: Spilker B (ed) Quality of life and pharmacoeconomics in clinical trials. Lippincott-Raven, Philadelphia, pp 575–585

Agresti A (1996) An introduction to categorical data analysis. Wiley, New York

AMDP&CIPS (1990) Rating scales for psychiatry, European edn. Beltz Test GmbH, Weinheim

APA (American Psychiatric Association) (2013) Diagnostic and statistical manual of mental disorders: DSM-5,

5th edn. American Psychiatric Association, Washington, DC
APA (American Psychological Association) (2014) Standards for educational and psychological testing. American Educational Research Association, Washington, DC. http://www.apa.org/science/programs/testing/standards.aspx
Asadi-Lari M, Tamburini M, Gray D (2004) Patients' needs, satisfaction, and health related quality of life: towards a comprehensive model. Health Qual Life Outcomes 2:32
Biel V. Big data analytics. http://www.webopedia.com/TERM/B/big_data_analytics.html
Bland JM, Altman DG (1996) Measurement error and correlation coefficients. Br J Med 313:41–42
Borg I, Groenen P (2005) Modern multidimensional scaling: theory and applications, 2nd edn. Springer, New York. 2nd edn. Washington, DC: American Council on Education
Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
Cohen J (1977) Statistical power analysis for the behavioural sciences, revth edn. Academic, New York
Committee for Medicinal Products for Human Use (2005) Reflection paper on the regulatory guidance for the use of Health-Related Quality of Life (HRQL) measures in the evaluation of medicinal products. EMEA, London
Cronbach LJ (1971) Test validation. In: Thorndike RL (ed) Educational measurement. American Council on Education, Washington, DC, pp 443–507
Cronbach LJ (2004) My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 64:391–418
Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52:281–302
EMEA (2007) Innovative drug development approaches, EMEA/127318/2007. EMEA, London
Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York, pp 38–46
Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613–619
Food and Drug Administration (2000) Title 21 Code of Federal Regulations (21 CFR Part 11) Electronic Records; Electronic Signatures. http://www.fda.gov/ora/compliance_ref/Part11/
Food and Drug Administration (2006). Guidance for industry, patient-reported outcome measures: use in medical product development to support labeling claims, Silver Spring
Food and Drug Administration (2009) Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. Available from: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. Accessed 3 Jan 2017
Food and Drug Administration (2016) Clinical outcome assessment (COA): glossary of terms. Available from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm. Accessed 3 Jan 2017
Food and Drug Administration (2017) Clinical outcome assessment qualification program. https://www.fda.gov/drugs/developmentapprovalprocess/drugdevelopmenttoolsqualificationprogram/ucm284077.htm. Page Last Updated: 06 Aug 2017
Gartner IT. Glossary. https://research.gartner.com/definition-whatis-big-data?resId=3002918&srcId=1-8163325102
Guilford JP (1946) New standards for test evaluation. Educ Psychol Meas 6(5):427–439
Hamilton M (1960) A rating scale for depression. J Neurol Neurosurg Psychiatry 23:56–62
Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York
Kelly P (2015) Electronic questionnaires at statistics Canady. https://wwwn.cdc.gov/qbank/QUest/2015/s423-paul_kelly_quest2015.pdf
Kendell R, Jablensky A (2003) Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 160(1):4–12
Kundel HL, Nodine CF, Conant EF, Weinstein SP (2007) Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology 242:396–402
Lienert GA, Raatz V (1998) Testaufbau und Testanalyse. Weinheim, Psychologie Verlags Union 1998 (6. Aufl)
Linden WJ, van der Hambleton RK (1997) Handbook of modern item response theory. Springer, New York
Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York
Messick S (1989) Validity. In: Linn R (ed) Educational measurement, 3rd edn. American Council on Education and Macmillan, New York, pp 13–103
Middel B, van Sonderen E (2002) Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. Int J Integ Care 2.:2002:e15
NASA Report. ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf
National Institute of Mental Health. Research priorities. NIMH page on RDoC. http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml
Obuchowski NA (2005) ROC analysis. Am J Roentgenol 184:364–372
Osgood CE, Suci G, Tannenbaum P (1957) The measurement of meaning. University of Illinois Press, Urbana
Patrick DL et al (2007) Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 10(Supplement 2):S125–S137

Powers JH, Patrick DL, Walton MK, Marquis P, Cano S, Hobart J, Isaac M, Vamvakas S, Slagle A, Molsen E, Burke LB (2017) Clinician-reported outcome assessments of treatment benefit: report of the ISPOR clinical outcome assessment emerging good practices task force. Value Health 20(1):2–14. https://doi.org/10.1016/j.jval.2016.11.005

Rasch G (1960) Probabilistic models for some intelligence and achievement tests. Danish Institute for Educational Research (expanded edition), 1980, Copenhagen. University of Chicago Press, Chicago

Rost J (2004) Lehrbuch Testtheorie – Testkonstruktion. Huber, Bern (2. Aufl)

Schadow G, McDonald CJ (2014) The unified code for units of measure. Regenstrief Institute, Inc. and the UCUM Organization 1988–2014. http://unitsofmeasure.org/ucum.html

Schwarz N (1999) Self-reports. How questions shape the answer. Am Psychol 54(2):93–105

Stevens SS (1951) Mathematics, measurement and psychophysics. In: Stevens SS (ed) Handbook of experimental psychology. Wiley, New York, pp 1–49

Thompson B (ed) (2003) Score reliability: contemporary thinking on reliability issues. Sage, Thousand Oaks

Walton MK, Powers JH III, Hobart J et al (2015) Clinical outcome assessments: conceptual foundation: report of the ISPOR clinical outcomes assessment – emerging good practices for outcomes research task force. Value Health 18:741–752

WHO. http://www.who.int/classifications/icd/en/

Wild D, Eremenco S, Mear I, Martin M, Houchin C, Gawlicki M, Hareendran A, Wiklund I, Chong LY, von Maltzahn R, Cohen L, Molsen E (2009) Multinational trials – recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR patient-reported outcomes translation and linguistic validation good research practices task force report. Value Health 12(4.) 2009):430

Wilke RJ, Burke LB, Erickson P (2004) Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved labels. Control Clin Trials 25:535–552

William HL, Conway MA, Cohen G (2008) Autobiographical memory. In: Cohen G, Conway M (eds) Memory in the real world, 3rd edn. Psychology Press, Hove