1 Introduction

1.1 Why this textbook on measurement in medicine?

Measurements are central to clinical practice and medical and health research. They form the basis of diagnosis, prognosis and evaluation of the results of medical interventions. Advances in diagnosis and care that were made possible, for example, by the widespread use of the Apgar scale and various imaging techniques, show the power of well-designed, appropriate measures. The key words here are 'well-designed' and 'appropriate'. A decision-maker must know that the measure used is adequate for its purpose, how it compares with similar measures and how to interpret the results it produces.

For every patient or population group, there are numerous instruments that can be used to measure clinical condition or health status, and new ones are still being developed. However, in the abundance of available instruments, many have been poorly or insufficiently validated. This book primarily serves as a guide to evaluate properties of existing measurement instruments in medicine, enabling researchers and clinicians to avoid using poorly validated ones or alerting them to the need for further validation.

When many measurement instruments are available, we face the challenge of choosing the most appropriate one in a given situation. This is the second purpose of this book. Researchers need systematic methods to compare the content and measurement properties of instruments. This book provides guidelines for researchers as they appraise and compare content and measurement properties.

Thirdly, if there is no adequate measurement instrument available, a new one will have to be developed, and it should naturally be of high quality. We describe the practical steps involved in developing new measurement instruments, together with the theoretical background. We want to help

researchers who take the time and make the effort to develop an instrument that meets their specific needs.

Finally, evaluation of the quality of measurements is a core element of various scientific disciplines, such as psychometrics, epidemiology and biostatistics. Although methodology and terminology vary from discipline to discipline, their main objective is to assess and improve measurements. The fourth reason for this book is therefore to integrate knowledge from different disciplines, in order to provide researchers and clinicians with the best methods and ways to assess, appraise, and improve the methodological quality of their measurements.

1.2 Clinimetrics versus psychometrics

Psychometrics is a methodological discipline with its roots in psychological research. Within the field of psychometrics, various measurement theories have been generated, such as classical test theory and item response theory (Lord and Novick, 1968; Nunnally, 1978; Embretson and Reise, 2000). These theories will be further explained in Chapter 2. Cronbach and Spearman were two famous psychometricians. Psychometric methods are increasingly applied to other fields as well such as medicine and health.

The term 'clinimetrics' is indissolubly connected to Feinstein, who defined it as 'measurement of clinical phenomena'. He focused on the construction of clinical indexes, and promoted the use of clinical expertise, rather than statistical techniques, to develop measurement instruments (Feinstein, 1987).

However, in this book we avoid using the terms psychometrics and clinimetrics. Our basic viewpoint is that measurements in medicine should be performed using the most adequate methods. We do not label any of these as psychometric or clinimetric methods, but we do indicate which underlying theories, models and methods are applied.

1.3 Terminology and definitions

Literature on measurement can be confusing because of wide variation in names given to specific measurement properties and how they are defined. Often, many synonyms are used to identify the same measurement property.

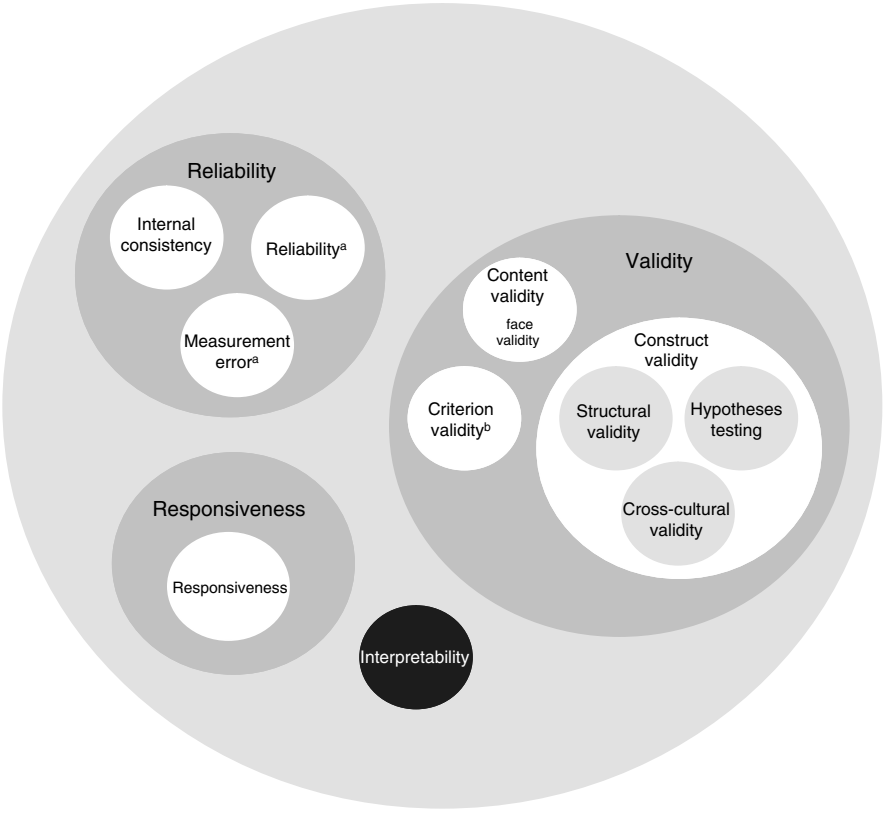
For example, the measurement property reliability is also referred to as reproducibility, stability, repeatability and precision. Moreover, different definitions are used for the same property. For example, there are many definitions of responsiveness, which results in the use of different methods to evaluate responsiveness, and this may consequently lead to different conclusions (Terwee *et al.*, 2003).

This variation in terminology and definitions was one of the reasons to start an international Delphi study to achieve consensus based standards for the selection of health measurement instruments (the COSMIN study) (Mokkink *et al.*, 2010a). The COSMIN study aimed to reach consensus among approximately 50 experts, with a background in psychometrics, epidemiology, statistics, and clinical medicine, about which measurement properties are considered to be important, their most adequate terms and definitions and how they should be assessed in terms of study design and statistical methods.

We adhere to the COSMIN terminology throughout this book. Figure 1.1 presents the COSMIN taxonomy, showing terms for various measurement properties and their inter-relationships. In chapters focusing on measurement properties, we indicate other terms used in the literature for the same properties, and also present the COSMIN definitions.

1.4 Scope of measurements in medicine

The field of medicine is extremely diverse. There are so many different diseases, and we all know that health is not just the absence of disease. The World Health Organization (WHO) officially defined health as 'a state of complete physical, mental, and social well-being, not merely the absence of disease or infirmity'. Evaluating the effects of treatment or monitoring the disease course includes assessment of disease stages, severity of complaints and health-related quality of life. To broaden the scope further, measurements do not only include all outcome measurements, but also measurements performed to arrive at the correct diagnosis and those done to assess disease prognosis. Measurements are performed in clinical practice and for research purposes. This broad scope is also expressed in the types of measurements. Measurements vary from questions asked about symptoms during history-taking, to physical examinations and tests,



[a](test–retest, inter-rater, intra-rater); [b](concurrent validity, predictive validity)

Figure 1.1 COSMIN taxonomy of relationships of measurement properties. Reprinted from Mokkink *et al.* (2010a), with permission from Elsevier.

laboratory tests, imaging techniques, self-report questionnaires, and so on. The methods described in this book apply to all measurements in the field of medicine.

1.5 For whom is this book written?

This book is for clinicians and researchers working in medical and health sciences. This includes those who want to develop or evaluate measurement instruments themselves, and those who want to read and interpret the literature on them, in order to select the most adequate ones.

We present the theoretical background for measurements and measurement properties, and we provide methods for evaluating and improving the quality of measurements in medicine and the health sciences.

A prerequisite for a correct understanding of all concepts and principles explained in this book is basic knowledge about study designs (i.e. cross-sectional and longitudinal), essentials of diagnostic testing and basic knowledge of biostatistics (i.e. familiarity with correlation coefficients, *t*-tests and analysis of variance).

This book is not directed at any specific clinical discipline and is applicable to all fields in medicine and health. As a consequence, the reader will not find a list of the best measurement instruments for paediatrics, cancer or dementia, etc., but a description of how measurement instruments should be developed, and how measurement properties should be assessed and can be improved.

1.6 Structure of the book

The book starts with introductory chapters focusing on measurement theories and models. In particular, Chapter 2 describes the essentials of the classical test theory and the item response theory. Chapter 3 describes the development of a measurement instrument.

Chapters 4–7 then focus on measurement properties. Each chapter describes the theoretical background of a measurement property, and shows how this property is assessed. The structure of a measurement instrument is discussed, and the principles of factor analysis and internal consistency are introduced in Chapter 4. Reliability and validity are presented in Chapters 5 and 6. In health care, changes in disease or health status over time are important, so responsiveness is discussed in Chapter 7.

Interpretation of the results of measurements deserves its own chapter. This aspect is often neglected, but is ultimately the main purpose of measurements. In Chapter 8 we discuss the interpretability of the scores and change scores on measurement instruments, paying special attention to minimal important changes within patients, and response shift.

Finally, Chapter 9 puts all the pieces together by describing how to perform a systematic review of measurement properties. This is a systematic review of the literature to identify instruments relevant for specific

measurement situations and to assess the quality of their measurement properties.

1.7 Examples, data sets, software and assignments

We use real examples from research or clinical practice and, where possible, provide data sets for these examples. To enable readers to practise with the data and to see whether they can reproduce the results, data sets and syntaxes can be found on the website www.clinimetrics.nl.

For statistical analyses, we used the Statistical Package for the Social Sciences (SPSS). For analyses that cannot be performed in SPSS, we suggest alternative programs.

Each chapter ends with assignments related to the theories and examples covered in that chapter. Solutions to these assignments can also be found on the website www.clinimetrics.nl.

2

Concepts, theories and models, and types of measurements

2.1 Introduction

This chapter forms the backbone of the book. It deals with choices and decisions about *what* we measure and *how* we measure it. In other words, this chapter deals with the conceptual model behind the content of the measurements (*what*), and the methods of measurements and theories on which these are based (*how*). As described in Chapter 1, the scope of measurement in medicine is broad and covers many and quite different concepts. It is essential to define explicitly what we want to measure, as that is the 'beginning of wisdom'.

In this chapter, we will introduce many new terms. An overview of these terms and their explanations is provided in Table 2.1.

Different concepts and constructs require different methods of measurement. This concerns not only the type of measurement instrument, for example an X-ray, performance test or questionnaire, but also the measurement theory underlying the measurements. Many of you may have heard of classical test theory (CTT), and some may also be familiar with item response theory (IRT). Both are measurement theories. We will explain the essentials of different measurement theories and discuss the assumptions to be made.

2.2 Conceptual models

First, we will look at the concepts to be measured. Wilson and Cleary (1995) presented a conceptual model for measuring the concept health-related quality of life (HRQL). Studying this model in detail will allow us to distinguish

Table 2.1 Overview of terms used in this chapter

Term	Explanation
Concept	Global definition and demarcation of the subject of measurement.
Construct	A well-defined and precisely demarcated subject of measurement. By psychologists used for unobservable characteristics, such as intelligence, depression or health-related quality of life.
Conceptual model	Theoretical model of how different constructs within a concept are related (e.g. the Wilson and Cleary[a] model of health status).
Conceptual framework	A model representing the relationships between the items and the construct to be measured (e.g. reflective or formative model).
Measurement theory	A theory about how the scores generated by items represent the construct to be measured (e.g. classical test theory or item response theory).
Method of measurement	Method of data collection or type of measurement instrument used (e.g. imaging techniques, biochemical analyses, performance tests, interviews).
Patient-reported outcomes	A measurement of any aspect of a patient's health status that comes directly from the patient, without interpretation of the patient's responses by a physician or anyone else.
Non-patient-reported outcome measurement instruments	All other types of measurement instruments (e.g. clinician-based reports, imaging techniques, biochemical analyses or performance-based tests).
Health-related quality of life	An individual's perception of how an illness and its treatment affect the physical, mental and social aspects of his or her life.

[a] See Figure 2.1.

different levels of clinical and health measurements (Figure 2.1). The levels range from the molecular and cellular level to the impact of health or disease on individuals in their environment and their quality of life (QOL), which represents the level of a patient within his or her social environment.

We illustrate this conceptual model, using diabetes mellitus type 2 as an example. On the left-hand side, the physiological disturbances in cells, tissues or organ systems are described. These may lead to symptoms that subsequently affect the functional status of the patient. For example, in patients with diabetes the production of the hormone insulin is disturbed, leading to high levels of glucose in the blood. The patient's symptoms are tiredness or thirst. In the later phases of diabetes, there may be complications, such as retinopathy, which affects the patient's vision. Patients with diabetes are

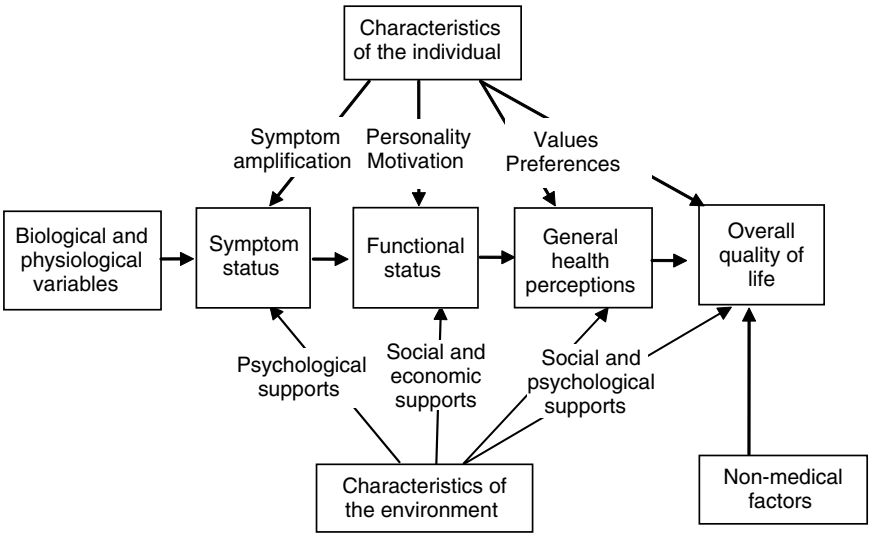


Figure 2.1 Relationships between measures of patient outcome in an HRQL conceptual model. Wilson and Cleary (1995), with permission. All rights reserved.

also more susceptible to depression. All these symptoms affect a patient's functioning. In the WHO definition of health, functioning encompasses all aspects of physical, psychological and social functioning. How patients perceive their health and how they deal with their limitations in functioning will depend on personal characteristics. Of course, the severity of the diabetes will affect the patient's functioning, but apart from that, a patient's coping behaviour is important. In addition, environmental characteristics play a role. For example, how demanding or stressful is the patient's job, and does the work situation allow the patient to adapt his or her activities to a new functional status? In HRQL, the factors we have described are integrated. Patients will weigh up all these aspects of their health status in their own way. Finally, in a patient's overall QOL, non-medical factors also play a role, such as financial situation or the country of residence. The Wilson and Cleary conceptual model illustrates how various aspects of health status are inter-related.

Wilson and Cleary developed their model not only to identify different levels of health, but also to hypothesize a causal pathway through which different factors influence HRQL. The arrows in the model indicate the most

important flows of influence, but Wilson and Cleary acknowledge that there may be reciprocal relationships. For example, patients with diabetes may become depressed because of their functional limitations and poor HRQL. Distinguishing different levels ranging from the cellular level to the societal level, looking from left to right in Figure 2.1, allows to focus on several measurement characteristics.

2.3 Characteristics of measurements

From diagnosis to outcome measurements

When diagnosing a disease, we often focus on the left-hand side of the Wilson and Cleary model, while for the evaluation of outcomes of disease or treatment the levels on the right-hand side are more relevant. The diagnosis of many diseases is based on morphological changes in tissues, disturbances in physiological processes, or pathophysiological findings. For example, a high blood glucose level is a specific indicator of diabetes because it reflects a dysfunction in insulin production. Other diseases, such as migraine and depression, can only be diagnosed by their symptoms.

Functional status is frequently considered an outcome of a disease. However, physiotherapists and rehabilitation physicians may consider it a diagnosis, because their treatment focuses on improvement of functioning. Further to the right in the model, perceived health and HRQL are typically outcome measures. None the less, disease outcomes can also be assessed by parameters on the left-hand side. For example, the effect of cancer therapies on the progression of cancer growth is usually evaluated on the basis of morphological or biochemical parameters at tissue level. At the same time, symptoms that bother patients and affect their HRQL are of interest. This example shows that the outcome of cancer is assessed at different levels, ranging from biological parameters to HRQL. However, diagnoses are usually found on the left-hand side of the model.

From clinician-based to patient-based measurements

Measurements performed either by clinicians or by patients themselves have different locations in the Wilson and Cleary model. Measurements of aspects on the left-hand side of Figure 2.1, either for the purpose of diagnosis or