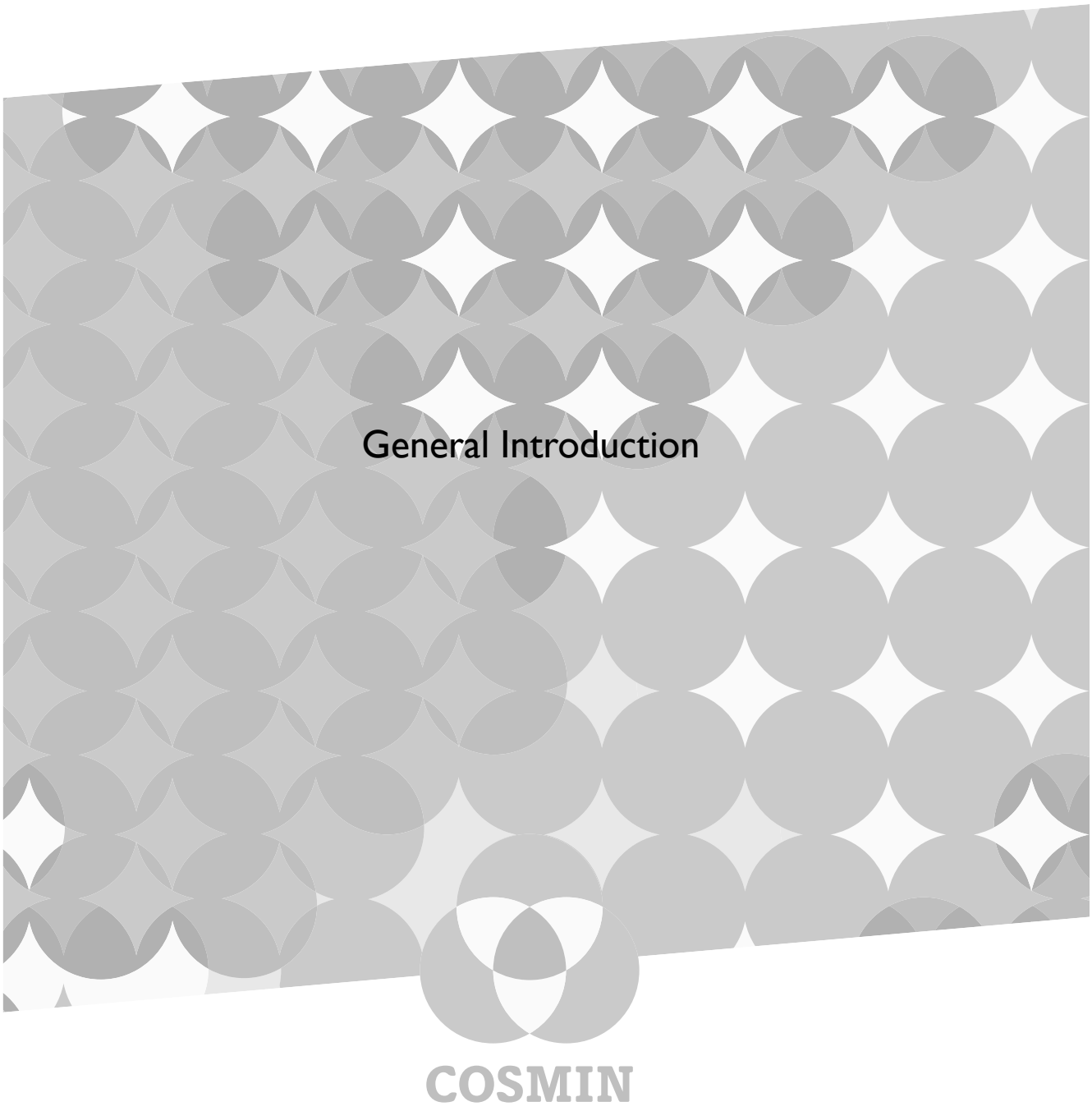
COSMIN: Development and evaluation of a checklist to assess the methodological quality of studies on measurement properties

Mokkink, L.B.

2010

document version
Publisher's PDF, also known as Version of record

I



General Introduction

The COSMIN initiative

COSMIN is an acronym for COnsensus-based Standards for the selection of health Measurement INstruments. The COSMIN initiative aims to improve the selection of health measurement instruments.

Within the medical field, there are many measurement instruments developed which aim to measure the same construct, even for the same patient population. Nevertheless, still new ones are developed. This has resulted in an overload of instruments. These instruments may not have equal quality. To evaluate the quality of a measurement instrument, its measurement properties (e.g. measurement error or content validity) need to be assessed.
The studies in which the measurement properties are being investigated should be methodologically sound. Studies of low methodological quality may give biased results, and consequently cannot be trusted. Therefore, the methodological quality of studies on measurement properties should be evaluated to detect potential bias.

This dissertation is about the COSMIN checklist, which can be used to assess the methodological quality of studies on one or more measurement properties. In this chapter, first some issues about quality assessment in general are discussed, followed by quality assessment specifically for studies on measurement properties. Next, the COSMIN Delphi study in which the checklist is developed will be discussed. This chapter will be closed with a description of the objectives and outline of the dissertation.

Quality assessment

Guidelines for design and performance

Assessing the methodological quality of studies is needed to detect potential bias in studies. Guidelines may be a useful tool for quality assessment. Methodological guidelines can help readers judging the quality of a study, i.e. whether the methods of the research correspond to the research question, and whether analyses, results and conclusions are also in line[1]. To assess the methodological quality of a study, a judgement about the appropriateness of the study design and the statistical methods used should be given. Well-known methodological guidelines are e.g. the Cochrane Collaboration's Tool for assessing risk of bias in randomized clinical trials (RCTs)[2], Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool for diagnostic studies[3], and the standards of the American Psychological Association (APA) for studies on measurement properties[4].

In this dissertation, the term "standards" will be used when talking about methodological guidelines for measurement properties. These standards can be used to give a judgement about the quality of a study. Standards contain requirements for study design and appropriate statistical methods.

Guidelines for reporting

To be able to adequately judge the methodological quality of a study, adequate reporting is required. Reporting guidelines for different types of studies have been developed. These guidelines contain statements that provide advice on how to report research methods and findings[5]. They aim to improve the accuracy and transparency of publications, thus facilitating easier and more reliable appraisal of quality and relevance[6]. Well-known reporting guidelines are e.g. the CONsolidated Standards of Reporting Trials (CONSORT) statements (for reporting RCTs)[7,8], STAndards for the Reporting of Diagnostic accuracy studies (STARD; for reporting diagnostic studies)[9], STrengthening the Reporting of OBservational studies in Epidemiology (STROBE; for reporting observational studies)[10], and Transparent Reporting of Evaluations with Nonrandomized Designs (TREND; for reporting non-randomized studies)[11]. No reporting guidelines are developed yet to assess the quality of papers on measurement properties.

Often the content of items in methodological guidelines and reporting guidelines are quite similar. For example, an item in a methodological guideline for RCTs is 'Was a method of randomization performed?'[12], while the item in the reporting guideline for RCTs is 'Was described how participants were allocated to interventions?'[7].

Guidelines for valuing the outcomes of the study

After assessing the methodological quality of the study and the quality of reporting, the outcomes of a study needs to be judged to value their quality. After a study is performed, we want to know whether the intervention is effective, or whether a test is useful. Therefore, for each type of study we need guidelines for interpretation of the results. In this dissertation we call these guidelines "criteria of adequacy". These criteria may be situation dependent, because they are dependent on the consequences associated with the outcome of the study. For example, to determine the quality of an intervention investigated in a randomized clinical trial (RCT), the question is 'how large should the difference be between the effect of intervention under study and e.g. the placebo intervention in order to implement the new intervention?'. The answer depends on e.g. the nature of an intervention (i.e. invasive or non-invasive), or costs.

To determine the quality of a diagnostic test, the question is 'how high should the sensitivity of the test be?' The answer depends e.g. on the consequences of missing people. For

example, for screening on breast cancer, which has good treatment options when detected in an early stage, one requires a high value for sensitivity. To determine the usefulness of a measurement instrument, we need criteria to judge its measurement properties. Again, these are situation dependent. For example, when an instrument is used as an outcome in a large RCT, the measurement error may be larger than when it will be used for individual patient monitoring.

These criteria refer to the quality of a measurement instrument, and contain guidelines for good measurement properties.

Quality assessment of studies on measurement properties

Several methodological guidelines for studies on measurement properties have previously been proposed, such as the attributes and criteria of the Scientific Advisory Committee of the Medical Outcomes Trust (SAC-MOS)[13], the standards of the American Psychological Association (APA)[4], Terwee Criteria[14], and EMPRO[15]. The SAC-MOS and the Terwee criteria focussed on health status measures, EMPRO on PROs, and the APA standards on educational and psychological testing. The APA standards present only methodological guidelines. The SAC-MOS criteria present standards separate from suggestions of criteria. For example, one of the items for reliability of the SAC-MOS guideline is 'information of test-retest reliability or inter-rater reliability [should be] based on ICC'[13]. In addition, SAC-MOS give suggestions for criteria of adequacy, e.g., 'Commonly accepted minimal standards for reliability coefficients are 0.70 for group comparisons and 0.90-0.95 for individual comparisons'[13]. Other guidelines are a combination of methodological guidelines and criteria of adequacy, such as those proposed by Terwee et al.[14], and EMPRO[15]. For example, Terwee et al. proposed a criterion for criterion validity: 'convincing arguments that gold standard is "gold" AND correlation with gold standard ≥ 0.70'[14].

The evaluation of measurement properties can be done in studies applying Classical Test Theory (CTT) or Item Response Theory (IRT). Guidelines for both applications should be developed. Some guidelines give brief standards for studies that apply IRT. The SAC-MOS criteria, for example, stated 'IRT and confirmatory factor analysis can be used to evaluate cross-cultural equivalence through examination of differential item functioning (DIF)'[13]. However, in none of the guidelines are items included about e.g. the IRT model used, the method of estimation, or about assumptions for estimating parameters. This information is needed for a good interpretation of the IRT parameters.

Fact is that none of the existing methodological guidelines are widely used in e.g. systematic reviews of measurement properties[16]. Moreover, most of the guidelines, except for the APA

standards, are rather brief, and therefore not useful for a systematic review of measurement properties with the aim to discriminate between studies with good/moderate/poor methodological quality. We believe that guidelines developed by consensus, and with a sufficient level of detail, are more likely to be used in the field. Sufficient detail may lead to a more user-friendly checklist. Furthermore, guidelines for studies that apply Item Response Theory (IRT) models should be included in the checklist, and they should be applicable for health-related patient-reported outcome (HR PRO) instruments. We started the COSMIN Delphi study to develop a checklist which fulfils all these requirements.

The COSMIN Delphi study

The aim of the COSMIN Delphi study was to develop a checklist containing standards and criteria to evaluate studies on measurement properties. To achieve consensus among different disciplines, a Delphi study was performed, involving a multidisciplinary, international group of researchers with relevant expertise.
We considered an international Delphi study a useful study design, because our aim concerned the agreement on terminology, and definitions of measurement properties, and subsequently on study design and preferred statistical methods. These issues cannot empirically be studied. By including many experts in the field, we aim for achieving wide acceptance of the checklist.

One can only decide on how to investigate something, when "something" is defined. Therefore, clear terms and definitions of measurement properties are indispensable to develop methodological guidelines for measurement properties. However, both terms and definitions are not consequently used in the literature[17,18]. Different use of terminology may lead to confusion about which measurement property is evaluated. Differences in definitions may lead to confusion about which concept the measurement property represents, and how it should be assessed. Therefore, in this Delphi study, firstly, we aimed to reach consensus on the terms and definitions of the measurement properties, and developed a taxonomy of the relationships of the measurement properties. Secondly, we aimed to reach consensus on standards of design issues and preferred statistical methods. Thirdly, we aimed to reach consensus on criteria of adequacy.

In this study we decided to restrict our focus, as we were not sure whether studies on measurement properties of all types of measurement instruments and all their applications would have the same standards and criteria. First, we choose to focus on HR-PROs. Second, we focused on measurement instruments used in an evaluative application.

A patient-reported outcome (PRO) is a measurement of any aspect of a patient's health status that is directly assessed by the patient, i.e. without the interpretation of the patient's responses by a physician or anyone else[19]. Modes of data-collection for PROs include interviewer-administered instruments, self-reported instruments, or computer-administered instruments[19]. With the restriction of *health-related outcomes* we exclude constructs such as quality of care, or patient satisfaction. Examples of HR-PROs are questionnaires assessing symptoms, functional status, and health-related quality of life. These are constructs which are not directly measurable. Because of the subjective nature of these constructs, it is very important to evaluate whether the measurement instruments measure these constructs in a valid and reliable way.

By adding the restriction that HR-PROs should be used in an evaluative application, i.e. used in a longitudinal design, we excluded measurement instruments which are (1) only used in discriminative applications, (2) only used for predictive purposes, such as diagnostic or screening instruments, or (3) only used as independent variable, such as a determinant, confounder or effect-modifier. Note that an instrument *is* not discriminative, evaluative or predictive, but it is *used* for discriminative, evaluative or discriminative purposes. For example, an instrument that is originally developed to assess cross-sectionally the health status of a population may subsequently be used in an evaluative application. The consequence of an evaluative application is that instruments should be responsive, i.e. they should be able to detect changes over time.

Objectives and outline of this dissertation

In this dissertation the COSMIN Delphi study is described, in which a checklist containing standards to evaluate the methodological quality of studies on measurement properties was developed and evaluated.

The research questions of the Delphi study were:

1. Which measurements properties should be included in the assessment of evaluative HR-PROs, and how should they be defined?
2. How should these measurement properties be assessed in terms of study design and statistical analysis? (i.e. standards)
3. Which criteria should be applied to define what good measurement properties are?

The study design of the COSMIN Delphi study is presented in Chapter 2. As input for the Delphi study, a systematic review of systematic reviews of measurement properties was conducted and described in Chapter 3. Its research questions were 'to appraises were the

quality of the review process, to describe how authors assess the methodological quality of primary studies of measurement properties, and to describe how authors evaluate results of the studies'. The results of the Delphi study are described in the Chapters 4 to 6. These chapters contain the consensus reached on the terminology and definitions of the measurement properties, and the development of a taxonomy of relationships of measurement properties (Chapter 4), the presentation of the COSMIN checklist, and the consensus reached on the standards (Chapter 5), and a description of discussions that the Delphi panel has had on several topics (Chapter 6).

Due to lack of time and complexity of the issues in the first two research questions, we could not study the third research question. Therefore, criteria of adequacy were not developed.

In Chapter 7 the inter-rater reliability of each item of the COSMIN checklist was investigated. The reliability of items is an indication of a proper application of the COSMIN checklist by potential users, i.e. whether different raters interpret and apply the items similarly.

This dissertation closes with a general discussion (Chapter 8) and a summary in English and Dutch.

Reference List

(1) Vandenbroucke JP. (2009). STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUOROM, REMARK... and CONSORT: for whom does the guideline toll? Journal of Clinical Epidemiology, 62, 594-596.

(2) Higgins J, Altman DG, on behalf of the Cochrane Statistical Method Group and the Cochrane Bias Method Group. (2008). Assessing risk of bias in included studies. In: Higgins J, Green S, eds. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell, 187-242.

(3) Whiting PF, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology, 3, 25.

(4) American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2009). Standards for educational and psychological testing. [Rev. ed.] ed. Washington, DC: American Educational Research Association.

(5) www.equator-network.org. Accessed August 7, 2009.

(6) Altman DG, Simera I, Hoey J, Moher D, Schulz K. (2008). EQUATOR: reporting guidelines for health research. Lancet, 371, 1149-1150.

(7) Moher D, Schulz KF, Altman DG. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet, 357, 1191-1194.

(8) Altman DG, Schulz KF, Moher D et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Annals of Internal Medicine, 134, 663-694

(9) Bossuyt PM, Reitsma JB, Bruns DE et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clinical Chemistry, 49, 1-6.

(10) Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet, 370, 1453-1457.

(11) Des Jarlais DC, Lyles C, Crepaz N. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American Journal of Public Health, 94, 361-366.

(12) Verhagen AP, De Vet HCW, De Bie RA et al. (1998). The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. Journal of Clinical Epidemiology, 51, 1235-1241.

(13) Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research, 11, 193-205.

(14) Terwee CB, Bot SD, De Boer MR et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34-42.

(15) Valderas JM, Ferrer M, Mendivil J et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value in Health, 11, 700-708.

(16) Mokkink LB, Terwee CB, Stratford PW et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Quality of Life Research, 18, 313-333.

(17) Streiner DL, Norman GR. (2006). "Precision" and "accuracy": two terms that are neither. Journal of Clinical Epidemiology, 59, 327-330.

(18) Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. (2003). On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Quality of Life Research, 12, 349-362.

(19) U.S.Dept of Health & Human Services FDA Center for Drug Evaluation & Research, U.S.Dept of Health & Human Services FDA Center for Biologics Evaluation & Research, U.S.Dept of Health & Human Services FDA Center for Devices & Radiological Health. (2006). Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health and Quality of Life Outcomes, 4, 79.