The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes

Mokkink, L.B.; Terwee, C.B.; Patrick, D.L.; Alonso, J.; Stratford, P.W.; Knol, D.L.; Bouter, L.M.; de Vet, H.C.W.

published in
Journal of Clinical Epidemiology
2010

DOI (link to publisher)
10.1016/j.jclinepi.2010.02.006

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

ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 737−745

Journal of Clinical Epidemiology

The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes

Lidwine B. Mokkink[a,*], Caroline B. Terwee[a], Donald L. Patrick[b], Jordi Alonso[c,d], Paul W. Stratford[e,f], Dirk L. Knol[a], Lex M. Bouter[a,g], Henrica C.W. de Vet[a]

[a]*Department of Epidemiology and Biostatistics, The EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands*

[b]*Department of Health Services, University of Washington, Seattle, WA, USA*

[c]*Health Services Research Unit, Institut Municipal d'Investigacio Medica (IMIM-Hospital del Mar), Barcelona, Spain*

[d]*CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain*

[e]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada*

[f]*School of Rehabilitation Science, McMaster University, Hamilton, Canada*

[g]*Executive Board of VU University Amsterdam, Amsterdam, The Netherlands*

Accepted 5 February 2010

Abstract

Objective: Lack of consensus on taxonomy, terminology, and definitions has led to confusion about which measurement properties are relevant and which concepts they represent. The aim was to clarify and standardize terminology and definitions of measurement properties by reaching consensus among a group of experts and to develop a taxonomy of measurement properties relevant for evaluating health instruments.

Study Design and Setting: An international Delphi study with four written rounds was performed. Participating experts had a background in epidemiology, statistics, psychology, and clinical medicine. The panel was asked to rate their (dis)agreement about proposals on a five-point scale. Consensus was considered to be reached when at least 67% of the panel agreed.

Results: Of 91 invited experts, 57 agreed to participate and 43 actually participated. Consensus was reached on positions of measurement properties in the taxonomy (68−84%), terminology (74−88%, except for structural validity [56%]), and definitions of measurement properties (68−88%). The panel extensively discussed the positions of internal consistency and responsiveness in the taxonomy, the terms ''reliability'' and ''structural validity,'' and the definitions of internal consistency and reliability.

Conclusions: Consensus on taxonomy, terminology, and definitions of measurement properties was reached. Hopefully, this will lead to a more uniform use of terms and definitions in the literature on measurement properties.

Keywords: Delphi technique; Outcome assessment; Psychometrics; Quality of life; Questionnaire; Terminology; Classification

1. Introduction

A lack of consensus exists about terminology (how do we call it?) and definitions (what does it mean?) of measurement properties (such as reliability and validity) across the different fields that contribute to health measurement. For example, in the literature, many different terms for the same measurement property ''reliability'' are used interchangeably, such as reproducibility, reliability, repeatability, agreement, precision, variability, consistency, and stability [1]. At the same time, the term ''agreement'' is also used to indicate another measurement property, that is, ''measurement error.'' Different uses of terminology can lead to confusion about which measurement property is assessed. Differences in definitions may lead to confusion about which concept the measurement property represents and how it should be assessed. For example, responsiveness may be defined as ''the ability to detect clinically important change'' or as ''the ability to detect change in the construct to be measured.'' These definitions reflect different constructs. The choice of a definition leads to

* Corresponding author. Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. Tel.: +31-20-444-3819; fax: +31-20-444-6775.

E-mail address: w.mokkink@vumc.nl or http://www.emgo.nl or http://www.cosmin.nl (L.B. Mokkink).

0895-4356/$ − see front matter
doi: 10.1016/j.jclinepi.2010.02.006

What is new?

- International consensus on terminology and definitions of measurement properties.
- Development of a taxonomy of the relationships of measurement properties.
- Multidisciplinary international collaboration leading to consensus.

different ways of assessing a measurement property. For example, Terwee et al. [2] calculated a number of parameters for responsiveness on the same data set, for example, an effect size (ES) (as a parameter to detect clinically important change) and a receiver operator characteristic (ROC) curve (as a parameter to detect change in the construct being measured). They found an ES of 0.39, which can be considered as moderate and an ROC curve of 0.47, which is poor [2]. Thus, the result and the conclusion of a study on measurement properties are dependent on the parameter used. Therefore, using one definition and its corresponding parameter may lead to other results and conclusions than using another definition and parameter.

A taxonomy of the relationships among measurement properties provides a complete picture of the relevant measurement properties when assessing the quality of health-related patient-reported outcomes (HR-PROs). A taxonomy is a classification containing domains and subcategories. In our taxonomy, the measurement properties and aspects of measurement properties are the subcategories.

To improve the field of assessing measurement properties, it is of utmost importance to reach consensus about terminology, definitions, and a taxonomy of the relationships of measurement properties of HR-PROs. A Delphi procedure is an appropriate design to reach consensus among experts [3,4].

The aim of this article is to clarify and standardize terminology and definitions of measurement properties by reaching consensus among a group of experts. In addition, the aim was to develop a taxonomy of the relationships of relevant measurement properties for evaluating HR-PRO instruments.

This study is part of the COSMIN initiative (COnsensus-based Standards for the selection of health Measurement INstruments), which aims to improve the selection of health measurement instruments. Within the COSMIN initiative, we performed a Delphi study in which we aimed to reach consensus on (1) which measurement properties are relevant for evaluating HR-PROs; (2) terminology and definitions of these measurement properties; and (3) the design requirements and preferred statistical methods. These issues should all be in line with each other to avoid confusion. Consensus reached on these issues is needed so that more uniformity is obtained in the use of terminology, definitions, and subsequently the design requirements and statistical methods. Results of studies on measurement properties are then better comparable. Moreover, consensus on these issues can lead to more understanding about what important measurement properties of measurement instruments are and how they should be investigated. In a related article, we describe the results of the Delphi study in which the panel reached consensus on which measurement properties are relevant for evaluating HR-PROs [5]. These are internal consistency, reliability, measurement error, content validity (including face validity), construct validity (subdivided into structural validity, hypotheses testing, and cross-cultural validity), criterion validity, and responsiveness. Interpretability was also considered to be relevant, although it was not considered to be a measurement property. In the related article, we also describe the consensus reached on design requirements and preferred statistical methods.

2. Methods

2.1. Steering Committee

A Steering Committee was formed to initiate and guide the study. Members of the Steering Committee are the authors of this article: five epidemiologists (L.B.M., C.B.T., D.L.P., L.M.B., and H.C.W.d.V.), a clinician (J.A.), a physiotherapist (P.W.S.), and a psychometrician (D.L.K.). The Steering Committee was responsible for the selection of the panel members, the design of the questionnaires, the analysis of the responses, and the formulation of the feedback reports. The members of the Steering Committee did not take part as panel members.

2.2. Preparation for the Delphi study

The Steering Committee systematically searched the literature to find systematic reviews on measurement properties of health status measurement instruments [1] and methodological articles and textbooks that provided terminology and definitions of measurement properties. The extracted information about terminology and definitions was used as input for the items in the Delphi questionnaires. We considered all names provided to indicate measurement properties as *terms*. *Definition* was defined as an explanation of what a measurement property means. In this study, a *measurement property* is defined as "a feature of a measurement instrument that reflects the quality of the measurement instrument."

2.3. Selection of panel members

Based on previous experiences with Delphi studies [6,7], we decided to invite at least 80 experts to participate in our panel to ensure 30 responders in the last round. We aimed to include experts in the field of psychology, epidemiology,

statistics, and clinical medicine. Among those invited were authors of reviews of measurement properties and methodological articles or textbooks on measuring health. We also invited experts working in organizations focusing on the health measurement, such as the International Society for Quality of Life (ISOQOL), the MAPI Research Institute, the Cochrane PRO Methods Group, and the European Research Group on Health Outcomes (ERGHO). Experts were invited when they had at least five publications on health measurements in PubMed. People from different parts of the world were invited. Panel members remained anonymous for both the Steering Committee and the panel members, except for L.B.M., until after the fourth round.

2.4. Focus of the COSMIN checklist

We focused on HR-PROs used in an evaluative application. We chose to focus on HR-PROs because of the complexity of these instruments. These instruments are aimed to measure not directly measurable multidimensional constructs. The addition of evaluative application was necessary because it influences the relevance of some measurement properties. For example, a discriminatively used instrument should be reliable and valid but not necessarily responsive. An instrument used in an evaluative application should be reliable, valid, as well as responsive.

2.5. Procedure of the Delphi rounds

The Delphi study consisted of four written rounds. Consensus was reached on which measurement properties should be included when evaluating HR-PROs. The results of this consensus procedure are described elsewhere [5]. For each measurement property, we asked which term and definition the panel members considered to be the best. We first provided a list of alternatives, which we had extracted from the systematic reviews and the methodological literature. The panel members were asked to indicate their preferred option or suggest another term or definition. For example, for the term of the measurement property that we now call responsiveness, panel members could choose from ''responsiveness,'' ''sensitivity to change,'' ''longitudinal validity,'' ''longitudinal construct validity,'' ''internal responsiveness,'' ''external responsiveness,'' ''precision,'' or ''other'' where they could give an alternative term. In the second round, the most frequently chosen option for each of the terms and definitions was proposed, and the panel members were asked to rate their (dis)agreement on a five-point scale (strongly disagree—disagree—no opinion—agree—strongly agree).

We asked the panel whether or not they agreed with the position of each measurement property in the taxonomy. The first proposal of the taxonomy was introduced in the second round. For example, we asked the panel ''Do you consider responsiveness as: (1) a part of validity; (2) a separate measurement property; or (3) not relevant.'' Agreement was rated on the same five-point scale. The panel members could propose alternative positions of the measurement properties in the taxonomy.

Each round consisted of a mailed questionnaire. In rounds 2–4, a feedback report of the previous round was included. Questions were based on comments and responses of the panel members from the previous rounds and on input from the literature. The feedback report contained all results of the previous round, including arguments provided by the panel members.

2.6. Consensus

Consensus was considered to be reached when the rating of at least 67% of the panel members indicated ''agree'' or ''strongly agree'' on the five-point scale. If less than 67% agreement was reached for a question, we asked it again in the next round, providing pro and contra arguments given by the panel members, or we proposed an alternative. If no consensus was reached, the Steering Committee took the final decision.

3. Results

3.1. Panel members

We invited 91 experts to participate: 57 (63%) agreed to participate and 15 (16%) were unwilling or unable to participate. The main reason for nonparticipation was lack of time. Nineteen experts (21%) did not respond. The mean number (minimum–maximum) of years of experience that the panel members had in research on measuring health or comparable fields (e.g., in educational or psychological measurements) was 20 (6–40) years. Most of the panel members came from Northern America ($n = 25$) and Europe ($n = 29$), whereas two were from Australia and one was from Asia.

A flowchart of the participation rates per round is presented in Fig. 1 (ranging from 48% to 74%). Twenty panel members (35%) participated in all four rounds and eight panel members (14%) did not participate in any of the rounds. Six panel members (11%) dropped out during the process, and the main reason was lack of time. One of these panel members decided to withdraw after the second round because of strongly divergent opinions. The names of all panel members who completed at least one round can be found in the Acknowledgments section.

3.2. Consensus

We reached consensus on the large majority of issues. Table 1 presents the percentages of panel members who agreed on position in the taxonomy, terminology, and definitions of the measurement properties. It took longer to reach the consensus for some properties than for other properties. In the following, we describe the issues that evoked most discussions (percentage agreement between parentheses).

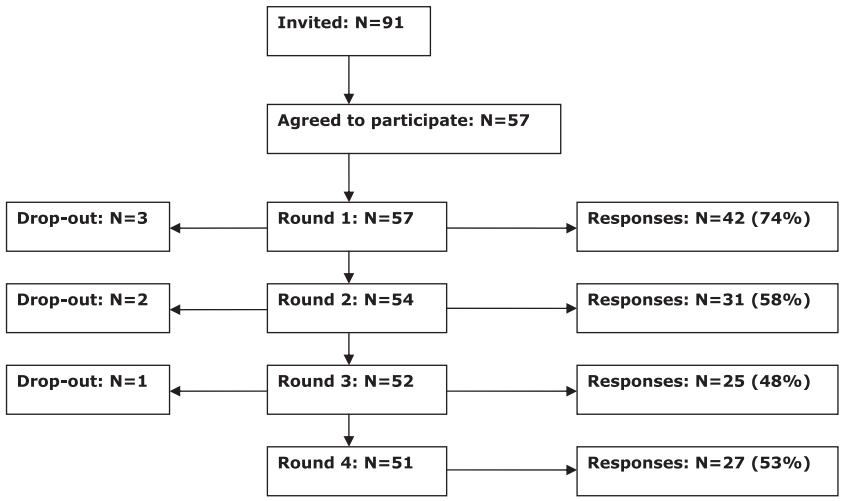

Fig. 1. Flowchart of participation rates per round.

3.3. Taxonomy

The taxonomy of the measurement properties is presented in Fig. 2. In assessing the quality of a HR-PRO instrument, we distinguish three quality domains, that is, reliability, validity, and responsiveness. Each domain contains one or more subcategories, that is, measurement properties or aspects of measurement properties. The domain reliability contains three measurement properties: internal consistency, reliability, and measurement error. The domain validity also contains three measurement properties: content validity, construct validity, and criterion validity. The domain responsiveness contains only one measurement property, which is also called responsiveness. The term and definition of the domain and measurement property responsiveness are actually the same, but they are distinguished in the taxonomy for reasons of clarity. Some measurement properties contain one or more aspects that were defined separately: content validity includes face validity, and construct validity includes structural validity, hypotheses testing, and cross-cultural validity.

The domain reliability contains internal consistency, reliability, and measurement error. We had a major discussion about the position of the measurement property internal consistency in the taxonomy. It took us to round 4 to reach consensus on this issue (81%). There was agreement that internal consistency belonged to the domain reliability. The discussion was on whether or not internal consistency should be considered a separate measurement property or an aspect of the measurement property reliability. This would be justified because of its similarity in way of calculation. However, for reasons of clarity and conceptual differences between internal consistency and reliability, the panel decided to consider it a separate measurement property (Fig. 2).

Table 1
Percentages of panel members who (strongly) agreed with the proposed position in the taxonomy, terminology, and definition

Domain	Measurement property	Aspect of a measurement property	Position in taxonomy	Terminology	Definition
Reliability			Nd	84 (R1)	77 (R2)
	Internal consistency		81 (R4)	84 (R2)	80 (R3)
	Reliability		84 (R2)	88 (R3)	70 (R4)
	Measurement error		74 (R2)	74 (R1)	87 (R2)
Validity			Nd	Nd	92 (R3)
	Content validity		77 (R2)[a]	Nd	81 (R2)
		Face validity	Nd	Nd	71 (R2)
	Construct validity		77 (R2)[a]	Nd	80 (R3)
		Structural validity	76 (R3)[b]	56 (DS)	68 (R2)
		Hypotheses testing	76 (R3)[b]	Nd	Nd
		Cross-cultural validity	76 (R3)[b]	74 (R2)	88 (R3)
	Criterion validity		77 (R2)[a]	Nd	71 (R2)
Responsiveness			68 (R3)	74 (R1)	74 (R3)
	Responsiveness		Idem[c]	Idem	Idem

Abbreviations: R, round in which consensus was reached; Nd, not discussed; DS, decision by the Steering Committee.
[a] % consensus to keep three forms of validity, that is, content validity, construct validity, and criterion validity.
[b] Consensus to distinguish between three types of construct validity, that is, structural validity, hypotheses testing, and cross-cultural validity.
[c] idem same as above.

Fig. 2. COSMIN taxonomy of relationships of measurement properties. *Abbreviations:* COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; HR-PRO, health related-patient reported outcome.

In the domain validity, we distinguished between the measurement properties content validity, construct validity, and criterion validity. Although important authors in the field of psychology have called all types of validity construct validity [8–10], we explicitly decided not to do so (77% agreement) because at the level of design and methods, these three forms of validity differ. Furthermore, we distinguished three aspects of construct validity, that is, structural validity, hypotheses testing, and cross-cultural validity (76% agreement). Because many different hypotheses can be tested which require various designs, hypotheses testing includes, for example, convergent, discriminant, and known groups validity. We explicitly distinguish these three aspects of construct validity for several reasons:

1. Structural validity should only be assessed for multi-item HR-PRO instruments, whereas the other aspects of construct validity are required for all HR-PRO instruments;
2. Structural validity should be assessed to determine or confirm existing subscales, which are subsequently used in the hypotheses that are being tested;
3. Cross-cultural validity should only be assessed for translated HR-PRO instruments.

We also had a major discussion about the position of the measurement property responsiveness in the taxonomy. In round 2, consensus was reached that the only difference between cross-sectional (construct and criterion) validity and responsiveness is that the first refers to the validity of a single score, and the latter refers to the validity of a change score (84% agreement). Sixty-eight percent of the panel members in round 3 were of the opinion that the entire domain responsiveness should be presented separately from the domain validity for reasons of clarity and to emphasis the distinction between the validity of a single score and the validity of the change score. Furthermore, we discussed whether the domain responsiveness should consist of two measurement properties, that is, construct responsiveness and criterion responsiveness, which is similar to construct validity and criterion validity. Panel members did not agree with this proposal because they did not want to introduce new terms.

3.4. Terminology

The term ‘‘reliability’’ is used twice, firstly as the term for the domain and secondly as the term for the measurement property (Fig. 2). In the first round, we reached consensus

on the term ''reliability'' for the domain (84%). For the term of the measurement property, we discussed whether we should use the term ''reliability'' or ''reproducibility.'' In round 1, 37% of the panel members chose ''reliability'' and 37% choose ''reproducibility'' for the term of this measurement property. Panel members argued that the term ''reproducibility'' has seldom been used for the measurement property that includes interrater reliability, intrarater reliability, and test–retest reliability. However, other panel members argued that the use of a different term, that is, reproducibility, may help to clarify that both this measurement property and internal consistency are forms of the domain reliability. Another panel member suggested that the term reproducibility might be easier to understand, whereas reliability seems to imply validity. In round 2, these arguments were reported back to the panel, and we asked the panel to choose between ''reliability'' and ''reproducibility.'' This time 55% of the panel members choose ''reproducibility'' and 32% choose ''reliability.'' Despite the higher percentage for ''reproducibility,'' the Steering Committee proposed to use the term ''reliability'' because at that time in the Delphi study, internal consistency was considered an aspect of this measurement property, and internal consistency had never been called ''reproducibility.'' In round 3, consensus was reached on the term reliability for this measurement property (88%). Note that the position of internal consistency was changed later and ended up as a separate measurement property.

The panel did not reach consensus on the choice between the terms structural validity and factorial validity: in round 2, 55% of the responding panel members agreed to use the term factorial validity, and in round 3, 56% agreed to use the term structural validity. Arguments against the term factorial validity were that it referred only to one of the methods to evaluate this aspect of construct validity, whereas structural validity referred to the purpose of this aspect. Based on these arguments, the Steering Committee decided to use the term structural validity.

The term for the domain validity as well as the terms for the measurement properties content validity, construct validity, criterion validity, and for the aspects face validity, concurrent validity, and predictive validity were not discussed in the Delphi study because in the literature there seemed to be no confusion about these terms.

3.5. Definitions

The definitions on which we reached consensus are presented in Table 2. We had a major discussion about the definition of internal consistency and in particular about the difference between internal consistency and homogeneity. Panel members proposed to use the articles written by Cronbach [8] and Cortina [11] to clearly distinguish between these two concepts. According to these sources, homogeneity of the items refers to the unidimensionality of a scale [11] and is a prerequisite for a clear interpretation of the internal consistency statistic [8,11]. Internal consistency is the interrelatedness among the items [11]. In round 2, the panel reached consensus on a definition of internal consistency (77%) in which we tried to combine both concepts, that is, ''the degree to which all items measure the same construct assuming the (sub)scale to be unidimensional.'' However, some panel members argued that the proposed definition should reflect what the statistics tell you, not infer anything beyond that. Based on this argument, we proposed Cortina's definition in round 3, that is, ''the interrelatedness among the items'' [11]. We reached 80% consensus on this definition.

We had another major discussion about the definition of the measurement property reliability. In round 2, the panel reached consensus on the definition (71%) ''the extent to which a measurement will give the same results on separate administrations.'' However, this definition is not in agreement with the most preferred statistical methods, that is, intraclass correlation coefficient or Cohen's kappa. These methods indicate the ability to differentiate among patients. Therefore, the Steering Committee proposed the (more statistical) definition of Streiner and Norman [12], that is, ''the proportion of the total variance in the measurements, which is because of ''true'' differences between patients.'' However, the panel members seemed to have problems with the word ''true,'' and only 56% agreed with this definition. Because of the ambiguous meaning of the word ''true'' in this context, an explanation was added: ''the word ''true'' must be seen in the context of the classical test theory, which states that any observation is composed of two components—a true score and error associated with the observation. ''True'' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score and not to its accuracy'' [12]. In the last round, 70% of the panel members agreed with the definition of reliability in combination with the explanation of the word ''true.''

In round 2, consensus was reached on the definition of responsiveness (we used the same definition both for the domain and for the measurement property because they are only distinguished for reasons of clarity) (74%) ''the ability of an instrument to detect important change over time in the construct to be measured.'' Based on suggestions made by the panel members in round 3, we proposed to remove the word ''important'' because the importance of the detected change is a separate issue that refers to the interpretation of the change score. In addition, the cutoff point between important change and nonimportant change is quite arbitrary. All the panel members who responded in round 3 agreed to this proposal.

4. Discussion

In an international Delphi study, consensus was reached on the positions of all relevant measurement properties of

Table 2
Definitions of domains, measurement properties, and aspects of measurement properties

Term			
Domain	Measurement property	Aspect of a measurement property	Definition
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: for example, using different sets of items from the same HR-PROs (internal consistency), over time (test–retest) by different persons on the same occasion (interrater) or by the same persons (i.e., raters or responders) on different occasions (intrarater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is because of "true"[a] differences among patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an HR-PRO instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the HR-PRO instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
	Criterion validity		The degree to which the scores of an HR-PRO instrument are an adequate reflection of a "gold standard"
Responsiveness			The ability of an HR-PRO instrument to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability[b]			The degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to an instrument's quantitative scores or change in scores.

Abbreviations: HR-PROs, health-related patient-reported outcomes; CTT, classical test theory.

[a] The word "true" must be seen in the context of the CTT, which states that any observation is composed of two components—a true score and error associated with the observation. "True" is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score and not to its accuracy (ref Streiner & Norman [12]).

[b] Interpretability is not considered a measurement property but an important characteristic of a measurement instrument.

HR-PRO instruments in a taxonomy. In addition, we reached consensus on terminology (percentage consensus ranging from 74% to 88% except for one term [only 56% agreed on structural validity]) and on definitions (percentage consensus ranging from 68% to 88%) for each included measurement property. Our aim of reaching consensus was to provide clarification and standardization on these complex issues hopefully leading to more uniformity in the evaluation of measurement properties.

The advantage of the taxonomy we developed is that relationships are explained between measurement properties that were considered relevant when investigating the quality of a measurement instrument. It shows which measurement properties belong together. For example, we consider internal consistency as part of the domain reliability. Sometimes, however, this is considered to be a measure for validity. Cortina [11] showed that it is not appropriate to consider this a measure of validity because it does not

tell you whether the items measure the construct that is intended to be measured. It only tells you whether the instrument measures something consistently but what that is would still be unknown [11].

Several initiatives were launched previously that proposed standards for assessing measurement properties. Well-known lists are 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) [14], and the quality criteria proposed by Terwee et al. [15]. COSMIN differs from these initiatives. The APA standards focus on educational and psychological testing, whereas the standards of the SAC-MOS, Terwee et al. [15] and COSMIN focus on health status measurement. The SAC-MOS standards and the Terwee criteria, however, are not based on consensus among a large group of experts [13,15]. Terminology and definitions used in these initiatives slightly differ from COSMIN. For example, APA does not include the domain of responsiveness [14]. In psychology, measures may be used more often in discriminative purposes in which responsiveness is not important. In health measurements, instruments are often used to evaluate patients over time, and in that situation responsiveness is an important measurement property. Terwee et al. [15] consider internal consistency not as a subcategory of the domain reliability and defined it as ''the extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct.'' This definition seems to consider internal consistency and unidimensionality as the same constructs. The COSMIN panel decided that these are different constructs. The SAC-MOS do not include measurement error in their overview table of measurement properties, although they state that in Item Response Theory (IRT) applications, a plot showing standard error of measurement (SEM) over the range of the scale is useful [13].

Hence, the consensus procedure among psychometricians, epidemiologists, biostatisticians, and clinicians led to a different taxonomy of measurement properties relevant for health measurement instruments.

A Delphi approach is useful for situations in which there is a lack of empirical evidence or when there are strong differences of opinion. The objectives of this study could not be investigated empirically. The questions concern issues, which are a matter of taste and tradition (e.g., terminology) or a matter of theoretical knowledge and logical reasoning (e.g., regarding the concepts and definitions of the measurement properties). For example, to reach consensus on the definition of internal consistency, we first had to reach consensus on its concept. More specifically, we had to distinguish between the concepts of homogeneity and internal consistency. This was based on theoretical knowledge. Once we distinguished the two concepts, the definition was a logical consequence of the consensus-based concept. Another example is responsiveness. We first reached consensus that the only difference between (construct and criterion) validity and responsiveness is that the first refers to the validity of a single score, and the latter refers to the validity of a change score. A logical consequence of this agreement was that the domain responsiveness should have a similar structure in the taxonomy and a similar definition as the domain validity. However, because panel members were not in favor of introducing new terms, we did not made a distinction between ''construct responsiveness'' and ''criterion responsiveness'' in the taxonomy. This was a matter of taste of the panel.

In addition, decisions were made taking into account the meaning of, for example, terms and definitions in other contexts. For example, the term ''sensitivity of change'' was not chosen because it might lead to confusion with the concept of sensitivity used in diagnostic research.

Strong differences of opinion seemed to be because of differences in expertise and scientific backgrounds of the experts and traditions between and within professional fields. By providing feedback from previous rounds, the Delphi technique provides the advantage of a group process of building on the work and expertise of all panel members. This was also acknowledged by some of the panel members in their feedback to us afterward.

It should be noted that reaching consensus in a Delphi process neither means that the correct answer has been found [3] nor that a correct answer exists. The output is an expert group's opinion and should be interpreted as such [4]. Some of the issues we discussed, for example terminology, are semantic issues with no ''correct answer.'' Sometimes even errors can result from the Delphi technique. For example, in round 2, we reached consensus on a definition of the measurement property reliability, which, as we later acknowledged, was not in agreement with the preferred statistical methods to assess reliability. Therefore, we proposed another definition in round 3 (Table 2). The Steering Committee has tried to avoid other errors and inconsistencies.

5. Conclusion

Consensus was reached on the position of measurement properties of HR-PRO instrument instruments in a taxonomy, on terminology, and on definitions of measurement properties. Lack of consensus in the literature has lead to confusion about which measurement properties are relevant, which concepts they represent, and how to assess these measurement properties in terms of design requirements and preferred statistical methods. Hopefully, this will lead to a more uniform use of terms and definitions in the literature on measurement properties.

Acknowledgments

The authors are grateful to all the panel members who have participated in the COSMIN study: Neil Aaronson,

Linda Abetz, Elena Andresen, Dorcas Beaton, Martijn Berger, Giorgio Bertolotti, Monika Bullinger, David Cella, Joost Dekker, Dominique Dubois, Anne Evers, Diane Fairclough, David Feeny, Raymond Fitzpatrick, Andrew Garratt, Francis Guillemin, Dennis Hart, Graeme Hawthorne, Ron Hays, Elizabeth Juniper, Robert Kane, Donna Lamping, Marissa Lassere, Matthew Liang, Kathleen Lohr, Patrick Marquis, Chris McCarthy, Elaine McColl, Ian McDowell, Don Mellenbergh, Mauro Niero, Geoffrey Norman, Manoj Pandey, Luis Rajmil, Bryce Reeve, Dennis Revicki, Margaret Rothman, Mirjam Sprangers, David Streiner, Gerold Stucki, Giulio Vidotto, Sharon Wood-Dauphinee, and Albert Wu. This study was financially supported by the EMGO Institute, VU University Medical Center, Amsterdam and the Anna Foundation, Leiden, The Netherlands. The funding sources had no role in the study design, the data collection, the data analysis, the data interpretation, or the publication.

References

[1] Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res 2009;18:313−33.

[2] Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12: 349−62.

[3] Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud 2001;38:195−200.

[4] Powell C. The Delphi technique: myths and realities. J Adv Nurs 2003;41:376−82.

[5] Mokkink LB, Terwee C, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539−49.

[6] Evers S, Goossens M, de Vet HCW, Van Tulder MW, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. Int J Technol Assess Health Care 2005;21:240−5.

[7] Verhagen AP, de Vet HCW, De Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 1998;51:1235−41.

[8] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297−334.

[9] Anastasi A. Psychological testing. New York, NY: Macmillan; 1988.

[10] Messick S. Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 1995;50:741−9.

[11] Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78:98−104.

[12] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford, UK: University Press; 2008.

[13] Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193−205.

[14] American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.

[15] Terwee CB, Bot SD, De Boer MR, Van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60: 34−42.