2nd year:	**Module 6**
Module Title:	**Measurement Theory Practice (Clinimetrics)**
Number of Credits:	**8 EC**

Aim of the module
The course will concentrate on grounds for selecting health status instruments in scientific research, the methodology and techniques of designing and evaluation of health status instruments and the critical appraisal of reports on the clinimetric properties of health status instruments.

Overall synopsis
The module starts with the background of "Clinimetrics" and presents some theoretical models for the classification of health outcomes. Topics of evaluating reliability and validity will receive great emphasis, together with the critical appraisal of reports on clinimetric or psychometric instruments. More specifically, the course will guide the students through the principals of classical test theory (CTT) and modern test theory (Item Response Theory) as frameworks for the evaluation of the reliability ('unidimensionality') of multi-item scales.

Learning outcomes as a whole
The general focus of this module is understanding the concept of Clinimetrics, its historical background, and the purpose of measuring health outcomes. The student is able to select appropriate health measurement instruments for clinical research designs and is aware of threats to the reliability and validity of instruments and can critically appraise articles on the metric properties of health status instruments. The student can design and analyze studies aimed at the development and / or testing of the metric properties (reliability / validity) of health status instruments at an intermediate level.

Teaching and learning strategies
Each session consists of a lecture and a workgroup. Lectures are given to guide the student through the relevant study material; workgroups are designed to get practical skills in critical appraisal and calculations. The student is encouraged to read textbooks for deeper understanding.

Assessment strategies
Knowledge of Clinimetrics will be tested at the end of the module by an exam using open-ended questions and forced choice questions.

Session 1: Lecture - Reliability I: Background of multi-item scales

The following topics will be addressed:

- Clinical endpoints: from "events" to multi-item scales.
- Understanding the latent variable: Classical Test Theory measurement assumptions regarding unidimensionality of multi-item scales measuring a theoretical construct, path diagrams, covariance matrix and internal consistency reliability for multi-item scales. Assessment of test-retest reliability and response stability .
- Distinction between generic, disease specific and domain specific multi-item instruments.
- Distinction between surrogate, intermediate and clinical endpoints.
- Generally accepted theoretical outcome models (Disablement process /ICF model, QoL model) .

- Levels of measurement according to S. Stevens 1947, incomparibility of scale scores scores in the abscence of a unit of measurement on instruments measuring a similar construct, and interpretation of: percentile scores, z-scores, T-scores, and population norm scores

Work group

- Components of internal consistency using: inter-item and item-rest correlations, split-half reliability, item-variance and covariance, contribution of individual items to scale reliability, calculation and interpretation of coefficient alpha, calculation of the Standard error of Measurement (SEM)
- Evaluation of test-retest reliability and assessment of response stability
- Levels of measurement (ordinal, nominal, interval, ratio)
- Classifying clinimetric endpoints using different medical outcomes models: ICF, QoL model. Recognition of clinical, intermediate and surrogate endpoints.

Learning outcomes

The student can assess the unidimensionality of multi-item scales used to measure some latent variable ("construct") . The student is able to carry out some basic psychometric analyses to evaluate the reliability (unidimensionality) of multi-item scales. Understanding of the distinction between surrogate, intermediate and clinical endpoints. Knowledge of globally used medical outcomes models

Session 2: Lecture - Reliability II: Reproducibility and agreement

This lecture will focus on:

- Measures of reliability and reproducibility: (weighted) Kappa for agreement on categorical scales (e.g. , item scores), Intraclass Correlation coefficient models for reproducibility of sum scores/ measures from multi-item scales, graphical method of Bland & Altman, SEM.
- Types of reliability: intra-method and inter-method reliability, inter-observer and intra-observer reliability

Workgroup

- Apply knowledge in computer lab using R statistical software package . Hands-on analysis of a dataset containing the scores of patients on different multi-item instruments.
- Practicing of methods to assess reproducibility of multi-item scales. Assessment of floor/ceiling effects, computation and interpretation of measures for score agreement on nominal, ordinal and interval scale, and graphical methods to evaluate score agreement.

Learning outcomes

The student has a broader understanding of the concept of reliability and has developed a critical attitude towards reports evaluating reliability as a clinimetric property. The student is able to carry out some basic psychometric analyses to evaluate reproducibility of scores or measures.

Session 3: Lecture - Reliability III : Exploratory Factor Analysis (principal components analysis)

The following topics will be addressed:

- Exploratory factor analysis as a method to investigate the dimensional structure of an item pool that has been presented to repondents. EFA can be used to identify interrelationships among items and to group items that are part of unidimensional constructs without making "a priori" assumptions about relationships among factors.
- Types of factoring with the emphasis on principal component analysis, a widely used method for factor extraction. The goal of PCA is to explain as much as possible the variation in scores in a dataset with successive factoring untill no further meaningful variance is left. Other types of factoring addressed include Common factor analysis (a.k.a. principal factor analysis or principal axis factoring), which seeks the least number of factors to account for the common variance between a set of items.
- Terminology that may be used for the presentation of a EFA and their meaning: communality for an item, factor loadings, eigenvalue of a factor, and factor scores.
- Criteria to determine the number of factors to extract: Kaiser criterium, variance explained criteria, Scree plot and parallel analysis.
- The main rotation methods (Varimax, Oblimin)

Work group

- Apply knowledge in computer lab using R statistical software package. Hands-on analysis of a dataset containing the scores of respondents .
- Practicing of methods to explore the dimensionality of a dataset.

Learning outcomes

The student can assess the dimensionality of multi-item scales used to measure a latent variable. The student is able to carry out an EFA to evaluate dimensionality of multi-item scales in order to judge the structural validity of an instrument. Student can explain the advantages of EFA: 1) reduction of number of variables by combining variables into a single factor, 2) identification of groups of inter-related variables to see how they are related to each other. Student can explain disadvantages of EFA: sample dependence, differing factor solutions in different samples. More than one interpretation can be made of the same data factored the same way. Correlational nature of the analysis: contrary to item response theory analysis (see reliability IV), factor analysis cannot identify causality.

Session 4: Lecture - Reliability IV: Item Response Theory

This lecture aims to a basic understanding of IRT as an alternative approach to one-dimensional scaling:

- Theoretical and practical flaws of CTT.
- Introduction to modern test theory,(IRT): basic assumptions, concepts and properties item trace line"), independent scaling subjects and items on a common logit unit scale.
- Special case of IRT model: the (basic) Rasch model and its extensions.
- Item and ability level invariance ("Specific objectivity"): demonstration of adaptive testing, generalizability of item statistics to new patients.
- Basic procedures for estimating item difficulty and discrimination and testing fit of item responses to the IRT/Rasch model.
- IRT extension of the classical concept of reliability: item-and test information, Standard Error of the Estimate (SEE)
- Application of IRT models to evaluate the quality of item rating scales.
- Understanding of the item-person map for score interpretation and
- Knowledge of the concept 'item bias' (differential item functioning)

- Advantages and disadvantages of IRT/Rasch models

Work group

- Apply knowledge in computer lab using R statistical software package. Hands-on IRT analysis of a dataset containing the scores of respondents . Perform tests to evaluate item fit, perform a rating scale analyis, and tests for unidimensionality of the item set as a whole. Plot item and test information functions and person-item map.

Learning outcomes

The student understands the strength and flaws of CTT, the basic concepts and methods, and historical background of modern scaling methods (IRT). The student has knowledge about the basic assumptions of IRT models, procedures to estimate the difficulty of items, testing of fit to the Rasch model, rating scale analysis and understands the basic properties of Rasch modeled scale: strict one-dimensionality, item statistics can be extrapolated to other populations ("specific objectivity").

Session 5: Lecture - Validity of multi-item construct scales

The issues addressed in this lecture are:

- Introduction to validity, and main types of validity
- Recognition and correct classification of types of validity, face content construct (convergent and divergent validity), criterion (concurrent and predictive validity)
- Face, content validation methods: focus groups or expert panels.
- Construct validity: convergent and divergent correlation analysis, known-groups validation method, criterion validity (concurrent /predictive validity).
- Measures to quantify or determine validity: correlations (construct validation), effect sizes, (known-groups validation), diagnostic test measures e.g., sensitivity, specificity, AUC analysis to assess the association between test score and some external reference standard (criterion related validation).

Work group

Apply knowledge in computer lab using the R statistical software. Hand-on analysis of a dataset containing the scores of patients on different multi-item instruments. Computational asssessment of construct validity: convergent, divergent validity, known- group validation, ROC-curve methods (criterion validation).

Learning outcomes

The student understands the concept of validity, the various types of validation including the appropriate statistical methodsand has developed a critical attitude towards reports evaluating validity as a clinimetric property of multi item health status instruments. The student can design a validation study and can carry out basic analysis to compare the validity of multi-item instruments

Session 6: Lecture - Putting it all together

Guest lecture: developmend and international validation of PROMs and a practice test.

Work group

Use of the COSMIN framework to score a research paper on a multi-item instrument on its psychometric or clinimetric quality.