Doctorate in Psychological, Anthropological, Educational Sciences

Department of Psychology, University of Turin

Development of a Multidimensional Computerized Adaptive short form of the Multiple Sclerosis Quality of Life-54 (MCAT-MSQOL-54): an international collaborative project

PhD candidate

Andrea Giordano

Tutor

Professor Rosalba Rosato

Contents

3. Development of a Multidimensional Computerized Adaptive short form of the Multiple Sclerosis Quality of Life-54 (MSQOL-54-MCAT): an international collaborative project

1. The health-related quality of life

Preface

Interest in measuring subjective (health-related) quality of life (HRQOL) and its outcomes increased over the last three decades. This was primarily due to the shift from caring acute diseases to the management of chronic, complex diseases. Such shift originated from several factors such as ageing populations, health campaigns, higher standards of living, and health technology development. Further, clinicians and researchers are increasingly asked to provide evidence on cost-effectiveness of interventions which are typically evaluated against the costs of such interventions [Cano 2011].

In the present chapter, concepts and definitions associated to (HR)QOL will be presented, along with its brief historical and theoretical perspectives. Further, contexts of use of HRQOL instruments will be reported. Then, a general overview and some examples of the most widely-used instruments assessing generic, disease-specific, as well as specific aspects of QOL will be provided.

HRQOL instruments are generally multidimensional and consist of single and multi-items scales. The concepts underlying such instruments are hypothetical, and should be measured by means of latent variables. To do so, traditional and modern psychometrics could be employed. Approaches such as factor modelling, item response theory, and computerized adaptive testing will be shortly presented with some pertinent examples from the literature.

THE DEFINITION OF (HEALTH-RELATED) QUALITY OF LIFE

A patient-reported outcome (PRO) has been defined as *'any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else'* [US FDA 2009]. The PRO is usually assessed either in broad terms (e.g. sign of a disease, symptom severity) or longitudinally, as a change from a prior evaluation. Self-reported or interview instruments can be used to assess PROs.

While the above mentioned definition of PRO is clear, the term quality of life (QOL) is still vague. In 1948, the World Health Organization (WHO) defined health '*[…] a state of complete physical, mental, and social well-being, and not merely the absence of a disease*' [WHO 1948]. The latter definition recognized the importance of these three dimensions – physical, mental and social in the context of disease. Patrick and Erickson expanded the WHO definition of QOL, as follows: *'[It] encompasses the entire range of human experience, states, perceptions and spheres of thought concerning the life of an individual or a community. Both objective and subjective, QOL can include cultural, physical, psychological, interpersonal, spiritual, financial, political, temporal, and philosophical dimensions. QOL implies judgement of value places on experience of communities, groups such as families, or individuals'* [Patrick and Erikson 1993].

More recently, the US FDA (2009) defined the QOL as '*A general concept that implies an evaluation of the effect of all aspects of life on general well-being. Because this term implies the evaluation of nonhealth-related aspects of life, and because the term generally is accepted to mean what the patient thinks it is, it is too general and undefined to be considered appropriate for a medical product claim'.*

There are a lot of other possible definitions of QOL and health which generally stressed the importance of happiness and life satisfaction. With no internationally agreed upon definition, it was argued by scholars that people are usually acquainted with the term QOL, and do usually grasp its

different components. At the same time, it is acknowledged that the meaning of QOL may depend on people and context of use.

When designing clinical trials, it was suggested to include components or aspects of QOL which are influenced by the disease/treatment. It would be possible also to include indirect consequences of disease (e.g. economic impact). In this case, the term HRQOL is normally used. The US FDA (2009) defined the HRQOL as*: '[…] a multidomain concept that represents the patient's general perception of the effect of illness and treatment on physical, psychological, and social aspects of life. Claiming a statistical and meaningful improvement in HRQOL implies: (1) that all HRQOL domains that are important to interpreting change in how the clinical trial's population feels or functions as a result of the targeted disease and its treatment were measured; (2) that a general improvement was demonstrated; and (3) that no decrement was demonstrated in any domain'.*

Thus, there is continuing debate about the meaning of QOL and about what should be measured. Perhaps the simplest and most pragmatic view is that all of these concepts reflect issues that are of fundamental importance to patient well-being.

In general terms, the QOL usually includes several components such as physical functioning, general health, emotional functioning, cognitive functioning, role functioning, physical symptoms and toxicity, sexual functioning, social well-being, and existential issues [Fayers 2016].

Some instruments/questionnaires concerned with objective signs (e.g. toxicity), others with a single concept, others with different QOL concepts/dimensions associated to several concepts. In practice, it is nowadays recognized that some of the above-mentioned components have to be evaluated, and that QOL is a multidimensional construct. Thus, instruments can include many items, as well as 'single global questions' which can be used to assess overall QOL, even if the latter are considered too vague. Other characteristics of the QOL instruments are that are subjective and should be administered to patients.

Historical perspective

From an historical perspective, the Karnofsky performance scale firstly broadened the assessment of patients beyond physiological and clinical aspects, focusing on functional ability. Over the years, it has led to a number of other scales focusing on functional ability, physical functioning, and activities of daily living, such as the Barthel Index [Katz 1976]. Although these instruments are described as QOL questionnaires, they grasp only one aspect of it, and provide no representation of patient overall wellbeing and QOL.

In the late 1970s and 1980s, other instruments which evaluated the patient health status including physical functioning, impact of illness, physical and psychological symptoms, life satisfaction, were developed, such as the Sickness Impact Profile and the Nottingham Health Profile. Although these instruments are frequently described as QOL instruments, their authors neither designed nor claimed as QOL instruments.

Meanwhile, another method was developed to assess QOL, by using a visual analogue scale (VAS) to assess for example, subjective effects of drugs in cancer patients. VAS for mood, anxiety, pain, social activities are typically used in this context [Priestman 1976].

Another approach considered a single global QOL question, and correlates this item score to other outcome measures [Gough 1983].

There are a number of theoretical models which were developed in QOL field and have to be briefly mentioned here as provide the theoretical background for the instruments presented in the section below.

Among others, the Expectations model [Calman 1984] was the theoretical base for two well-established instruments, such as the Patient Generated Index (PGI) and the Schedule for Evaluation of Individual Quality of Life-Direct Weighting (SEIQOL-DW), briefly described in a dedicated section

below. Here, QOL is a measure of the difference between hopes and expectations of the individual. It refers to the difference between perceived goals and actual goals [Calman 1984].

The 'needs model' relates QOL to the ability and capacity of patients to satisfy human needs. When the QOL level is high, all the needs are fulfilled; when it is at lower levels, few needs are satisfied. The needs included in the model were status, self-esteem, love, self-identity, security, sleep, pain, and food [Hunt 1992].

In the 'impact of illness' model, emotional, social, occupational, and issues about family were included. In the 'existential' approach, a positive attitude to life is considered as well coping with the disease. Here, a patient perception of his/her QOL can be altered by influencing their beliefs or by helping him/her to cope better with the disease. The existential approach led to the inclusion of items such as pleasure of life and positive outlook on life.

Finally, patient preferences instruments usually include 'weights', mirroring the importance patients attach to specific dimensions. Different states and dimensions are compared against each other, to establish a ranking in terms of their value. Thus, utility measures can be derived, as for example the EQ-5D (for details, see below).

Contexts of use of QOL instruments

QOL instruments can be normally used in trials with therapeutic purpose to assess primarily the effect of treatment on QOL, or as a secondary outcome to evaluate the effects of such treatment [Fayers 2016].

Another important context of use is palliative care. In this context, enhancing QOL is vital to assess the effect of such intervention on QOL, together with the assessment of patient symptoms, as reported in the definition of palliative care by the WHO.

Other essential contexts of use of QOL instruments are care and rehabilitation, as well as research aiming to facilitate communication with patients [Fayers 2016].

Investigation of patient preferences has been briefly mentioned above. In this case, patient weights have to be assessed in order to see for example, whether patient experiences or perceptions regarding any treatment differ from those of the physician or other health professionals [Fayers 2016].

In health-care decision-making, QOL could be an indicator of a success of any therapy, or clinical benefits could be contrasted against QOL [Fayers 2016].

HEALTH-RELATED QUALITY OF LIFE INSTRUMENTS

GENERIC INSTRUMENTS

Generic instruments were developed in order to assess multiple aspects or components of QOL. They are also used to compare the impact disease could have across various patient groups or healthy people. Generic instruments were initially designed to be used in population surveys. Then, they were largely-used in clinical trials. They were normally called 'health status' measures, as they particularly focused on physical functioning. When interpreting study results, at the poorer health status usually corresponds poorer QOL.

An international initiative which needs to be mentioned here is the International Quality of Life Assessment (IQOLA) project. This project was launched in 1992, with the aim to translate, adapt, and validate the SF-36 questionnaire into different languages [Aaronson 1992].

Generic measures allow straight comparisons between several patient sub-groups, thus giving the possibility to change policy and research agenda across a range of diseases. Nonetheless, major limitations are that these measures are not able to investigate key aspects of endpoints affected by a particular disease, and are not sensitive to change occurring to treatment or over time [Patrick 1989].

Some of the most widely QOL generic instruments are briefly presented below.

Sickness Impact Profile

It consists of 136 items, and it assesses health status, as impact on behavior. It is applicable to various diseases [Bergner 1981], and it focuses on everyday activities. It includes 12 areas of dysfunction/subscales (sleep and rest [example item: *'I sleep or nap during the day'*], eating [*'I am eating special or different food'*], work [*'I am not working at all'*], home management [*'I am not doing heavy work around the hours'*], recreation and pastimes [*'I am going out for entertainment*

less'], ambulation [*'I do not walk at all'*], mobility ['*I stay within one room*'], body care and movement ['*I am very clumsy in body move*'], social interaction [*'I isolate myself as much as I can from the rest of the family'*], alertness behavior [*'I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things'*], emotional behavior [*'I laugh or cry suddenly'*], communication [*'I am having trouble writing or typing'*]), and no question on overall QOL is included. The total score is calculated summing up the 12 subscales.

Nottingham Health Profile

It evaluates physical distress, emotional, social - feelings and emotions rather than behaviors.

Originally developed by Hunt et al. (1981), its version 2 includes 38 items divided in 6 sections (i.e. physical mobility, sleep, emotional responses, pain, social isolation, and energy level). It is a profile based on 6 sections, with no single total score. It is used either in health care or other different settings, but not in clinical studies.

Medical Outcomes Study 36-Item Short Form (SF-36)

The SF-36 inventory evaluates the general health status [Ware 1993]. It is a widely-used and generic instrument, which is not peculiar to age, disease or treatment groups. It focuses on physical, emotional and social functioning, including 36 items divided in 8 domains.

Several versions (i.e. 12-, 8-, 6-items are available, Qualitymetric 2021). Two summary scores are calculated: mental health composite, and physical health composite.

The physical health includes the following domains: physical functioning [*'Climb one flight'*], role physical [*'Limited in kind'*], bodily pain [*'Pain interfere with enjoyment'*], and general health [*'Worried for life'*]. The mental health includes the following domains: social functioning [*'Social time'*], energy [*'Worn out'*], mental health [*'Down in dumps'*], and role emotional [*'Not careful'*].

Also, a general health transition question *'Compared to one year ago, how would you rate your general health now?'*; and a global question on perception of their health - *'In general, would you say your health is (excellent, very good, good, fair, poor)?'* are included. The period to which subjects should refer when responding is 4 weeks, with items having different response categories.

A possible drawback of the SF-36 is the physical functioning domain, in that it is unclear how people who do not participate in physical activities (f.e. *'walking more than a mile'*, or *'vigorous activities, running, etc.)* should respond. Further, physical functioning dimension focuses on items related to actions which apply to anyone, while others emphasize that items are hypothetical.

EuroQol (EQ-5D)

Developed by the EuroQol group (EuroQol 2021], it includes physical, mental, and social functioning [Brooks 1996]. It is a simple instrument and should be used together with other QOL instruments/scales. Widely-used in clinical trial setting and at the international level, it is available in different versions (i.e. 5-, 3-items) investigating self-care, mobility, pain, usual activities, and anxiety/depression. Scores are determined by taking into account patient preferences. It is used also in health economics studies.

Patient Generated Index and the Schedule for Evaluation of Individual Quality of Life-Direct Weighting (SEIQOL-DW)

The most extensively used individualized instruments are the Patient Generated Index (PGI) [Ruta 1994], and the SEIQOL-DW [Hickey 1996], the shortened form of the SEIQOL. Either the PGI or the SEIQOL-DW are based on an interview to gather data and let subject to openly propose areas, followed by a scoring and weighting process. The main difference between the two measures is that the PGI focuses on the effect of the illness on patient QOL, while the SEIQOL-DW investigates QOL

in broader terms. Main barriers to their use are the following: first, trained personnel are needed to administer the interviews, thus their administration is time and resource-consuming. Second, interviewees might reduce the individual's notion of the nominated QOL domains.

DISEASE-SPECIFIC INSTRUMENTS

Disease-specific instruments are designed and used to complement generic instruments, by focusing on QOL aspects that are of particular interest to patients with disease/condition of interest. Rather than advocate using only one or the other, the most typical recommendation is to complement a generic measure with disease-targeted items.

They are generally more sensitive to small differences and small changes over time, in comparison to the generic instruments, because they are selected to be particularly relevant to a given condition. In fact, as reported in the examples below, they could include a generic core in addition to items assessing symptoms and problems associated with a given condition, or only the latter. Also, they could be used in tandem to provide the strengths of both approaches. For example, a short generic instrument as the SF-12 could be used together with a disease specific one (less response burden) to assess different aspects of QOL.

A brief description of the most widely-used disease-specific instruments is presented below.

European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (EORTC QLQ)-C30

It is applied in oncology patients, and consists of 30 items [Aaronson 1993]. It is multidimensional, and it is divided in five functional scales (role, physical, social, emotional, cognitive), 3 symptom scales (fatigue, pain, nausea and vomiting), a general QOL scale, additional symptoms, and economic

impact. It is widely-used, translated in several languages, and is sensitive to change. Higher scores indicate better functioning. Cancer modules have been developed to integrate the core part.

Functional Assessment of Cancer Therapy–General (FACT-G)

As the QLQ-30 above, FACT-G has a modular approach. Specific measures are available also for other disease, like multiple sclerosis. The FACT-G includes 27 items covering four dimensions: functional well-being, emotional well-being, physical well-being, and social well-being. It focuses on feeling and concerns [Cella 1993].

Rotterdam Symptom Checklist (RSCL)

The RSCL includes 30 items, focusing on symptoms and side effects in cancer patients [de Haes 1990]. It was widely-used in the past, but no more nowadays.

Quality of Life in Epilepsy (QOLIE-89)

Developed by Devinsky et al. (1995), and based on SF-36, QOLIE-89 items derived from several sources. Thirty-one- and 10-item versions are available. It consists of 17 multi-item scales: emotional well-being, overall QOL, role limitations, energy, social isolation, social support, seizure/worry, attention/concentration, physical function, health discouragement, memory, language, and health perceptions. General scores and four composite scores (epilepsy, cognition, mental health, and physical health) are determined.

Specific aspects of quality of life

Finally, scholars have also developed instruments assessing specific aspects of QOL and they are particularly useful for chronic illness or advanced disease. These instruments can allow to determine

an overall assessment of a specific aspect of QOL, which can be suitable to several patient groups and treatments. A possible advantage is that such instruments give a more thorough evaluation. A drawback is that it is not possible to calculate differences between patient groups.

Among others, it is noteworthy to mention the following instruments: McGill Pain Questionnaire (MPQ) [Cleeland 1994]; Hospital Anxiety Depression Scale (HADS) [Zigmond 1983]; Multidimensional Fatigue Inventory (MFI) [Smets 1995]; Barthel Index of Disability (BI) [Mahoney 1965].

SINGLE- AND MULTI-ITEM SCALES

Instruments assessing QOL usually include many items/questions, even if some attempts have been made in the past to assess QOL using a single overall question [Fayers 2016].

Some instruments require items to be summed up to produce a single score. Others require items to be divided in subscales related to different QOL dimensions.

For example, in the QLQ-30 instrument, a single question about vomiting is included. It can be considered inaccurate to assess such symptom with a single question, but in order not to lengthen the questionnaire it was deemed appropriate to do so. To further explore the effect of vomiting symptom on QOL, it may be useful to add a specific questionnaire on this symptom, if available.

Specific psychological constructs like anxiety and depression are difficult to define, as they can refer to different theories, that patients cannot understand. In many cases, psychological concepts are not directly measurable. These concepts can constitute latent traits, factors, or constructs. These hypothetical concepts are measured with latent variables. In our context, QOL itself can be considered as a latent variable.

The so-called factors are at a lower level. Physical functioning, social functions are all factors that can be considered latent variables or indicators of QOL.

A single trait which underlies the data is named unidimensional. As QOL includes usually many factors, it is a multidimensional construct, generally assessed with distinct dimensions, using a combination of single-item and multi-item scales.

Single-item and multi-item scales are commonly used in the QOL context. Single-item scales are often used to calculate a global QOL score. A possible drawback is that their responses are unreliable, imprecise, and challenging to interpret, as they are ill-defined. In fact, it could be more

appropriate to ask multiple questions about many QOL aspects, rather a single question [Fayers 2016].

The multi-item scales include more aspects likely to be unidimensional constructs. For example, intelligence is the construct, which has been analyzed in its different components. The same applies to QOL which includes different dimensions. Having many items, the multi-item scales increase the scope of a scale, reduce random error measurements, and improve the reliability of an instrument [Nunnally 1978]. Also, the presence of many response categories, as well as of many items, improve the precision of the instruments, providing additional information on the latent variable. The use of computerized adaptive testing (CAT) may be pertinent in this context, as CAT selects items at each step until a given level of precision is attained [Wainer 2000].

TRADITIONAL VERSUS MODERN PSYCHOMETRICS

Traditional psychometrics is based on summated score of Likert-type items to yield a score that represents the degree to which the construct being measured is present in the respondent [Likert 1932; Likert 1934]. The central idea is that latent variables are considered as unobserved determinants, the common cause of a set of observed (manifest) variables/indicators and are responsible of their covariations. Therefore, a researcher who views the QOL as a latent variable assumes that QOL is the common cause of the responses to a set of distinct QOL items. Usually, the researcher should set up a statistical model, i.e. a formal structure that relates observed scores to the hypothesized latent variable, deduces empirical implications of the model, and assesses the adequacy of the model, by examining the goodness of fit with regards to empirical data based on a substantive theory. The fundamental properties to be considered here are assumptions about scaling, reliability, validity, and responsiveness. In QOL context, two international initiatives conducted in US have to be mentioned. First, the Health Insurance Experiment made use of

psychometric methods to produce reliable and valid instruments to assess modifications in health status of child and adult general populations [Brook 1979]. Second, the Medical Outcomes Study [Stewart 1989] made use of psychometric methods to successfully construct scales and collect data to assess health status in sick and elderly people.

In contrast to traditional psychometrics, models investigating item response, are named modern psychometrics.

Factor models

Historically, the conceptual framework of factor models originated with Spearman (1904), who developed factor analytic models for continuous variables in the field of intelligence testing, but important progresses have followed in the 20th century, the most remarkable being the conceptual framework of confirmatory factor analysis (CFA) in 1969 by Joreskog.

In an exploratory factor analysis model, a response to each item is governed by a set of latent variables, called 'common factors', and by a single factor that incorporates the specificity of the individual item (a source of systematic variability not shared with other items) and the accidental error. The coefficients expressing the influence of common factors on manifest variables are called loadings. Common factors are all related to each other (oblique solution), or unrelated (orthogonal solution), unique factors are orthogonal to each other and all variables (observed and latent) are standardized. The starting data matrix is a correlation matrix. In a confirmatory factor analysis model, the researcher has the possibility to impose some constraints on the parameters, for example he/she can constrain to zero some loadings, making sure that each item depends on a single factor, or he/she can constrain to zero some of the correlations between the factors. The starting matrix is typically a covariance matrix.

Exploratory factor analysis is usually employed when validating QOL/PRO instruments, whereas confirmatory factor analysis is used to confirm a factorial structure hypothesized a priori [Rust 2009].

Item response theory

Item response theory (IRT) is frequently referred to as modern test theory or latent trait analysis. It is a paradigm specifically designed for observed categorical data. IRT models can be generally classified in accordance with the type of the data (i.e. dichotomous vs. polytomous items), and the number of the underlying latent traits (i.e. unidimensional vs. multidimensional model).

As distinct from factor modeling, IRT is an item-level theory. In IRT, a latent trait/ability/proficiency, usually symbolized as θ, is posited to underlie the observed responses. IRT models assume a specific relationship between θ and the observed responses. A common example of a model for dichotomous data is provided by the Rasch model, in which estimated parameters represent the positions of subjects and items on a latent continuum. Using a logistic function, item difficulty is computed in relation to person's ability, both placed on such linear continuum [Lord 1968; Embretson 2000]. The model assumes unidimensionality (a single underlying dimension), and local independence (i.e. no dependency among the items, after accounting for the latent trait). The local dependence can potentially derive from items having a comparable stem, content, or presented sequentially.

Among others, the most common models for dichotomous items are the 1, 2, and 3 parameter logistic models.

Variations of IRT model, such as the Rating Scale Model [Andrich 1978a, Andrich 1978b]; Partial Credit Model (PCM [Masters 1982]; the Generalized Partial Credit Model [Muraki 1992; Muraki

1997]) as well as the Graded Response Model (GRM) [Samejima 1969; Samejima 1997] are also available for ordered responses.

The GRM is a logistic model for graded responses data developed by Samejima. By definition, the probability of responding by an examinee with a specific level of ability θ_i to the category k of an item j is determined by the difference between the cumulative probability of a response to that category or higher, and the cumulative probability of a response to the next highest category (k+1) or higher, as follows:

$$P_{ijk}(\theta) = P'_{ijk}(\theta_i) - P'_{ijk+1}(\theta_i)$$

$$P'_{ijk}(\theta) = \frac{1}{1 + \exp\left[Da_j\left(\theta_i - b_{jk}\right)\right]}$$

Here, b_{jk} corresponds to the difficulty for category k_j and a_j to the discrimination parameter for item *j* and *D* is a scaling factor [Samejima 1969].

Further, besides having polytomous items, in psychology and QOL fields questionnaires generally measure multidimensional latent traits, so it would be appropriate to use multidimensional (MIRT) models, which overcome the unidimensionality framework of IRT models, by investigating a more generalizable model to fit the data [Seo & Weiss 2015].

IRT underpins computer adaptive testing (CAT), in which it is assumed that items are interchangeable, in such a way that various responders are administered different subsets of items.

Computerized adaptive testing

The computerized adaptive testing (CAT) is nowadays a cornerstone of standardized testing over the world. Based on the early attempts and implementations by the US military agents, the CAT began seriously in 1980s when cheap, high-powered computers were made available. Computer-based delivery of tests can be employed to increase statistical accuracy of test scores using CAT. Rather than administering each respondent with the same fixed-test, CAT item selection adapts to the ability level of individual respondent. After each response the person's ability estimate is updated and the subsequent item is selected in order to have optimal properties at the new estimate.

The idea of adapting the selection of the items to the examinee is certainly not new. In the Binet-Simon (1905) intelligence test, the items were classified according to mental age, and the examiner was instructed to infer the mental age of the examinee from the earlier responses to the items and to adapt the selection of the subsequent items to his/her estimate until the correct age could be identified with sufficient certainty.

The development of IRT in the middle of the last century has provided a sound psychometric basis for CAT. As mentioned above, the key feature of IRT is its modeling of response behavior with distinct parameters for the respondent's ability and the characteristics of the items. Due to this parameter separation, the question of optimal item parameter values for the estimation of individual ability became relevant. In 1968 Birnbaum proved that, unless guessing is possible, the optimal item is the one with the highest value for the item discrimination parameter and a value for the difficulty parameter equal to the ability of the respondent. QOL/PRO instruments have usually a fixed-format, are paper-based, and typically all patients are administered all items. CAT allows to the items to be customized to the individual participant, maximizing the information obtained.

This procedure has several advantages: the CAT version of the questionnaires can be 30-50% shorter than the traditional (paper-based) version, and the scores be estimated with more precision. By connecting the IRT approach with strong computer capabilities, CAT presents an encouraging research area in QOL/PRO assessment. The starting point is typically an item bank including questions which were calibrated according to psychometric techniques [Revicki 1997; Wainer 2000]. An Item bank includes a huge number of items with various difficulty levels and covering different levels of ability. The items included in the Item bank originated from a literature review and different instruments. Further, it is also possible to create new items to complement the existing bank.

Considering a few examples of item banks, the FACIT group set up an item bank to evaluate cancer-related fatigue. Based on the FACIT-Fatigue subscale, the authors used IRT to calibrate items. Findings showed that a 4-item fatigue scale was equivalent to the original 13-item version [Lai 2003]. Another example derives from the EORTC QOL group. Here, Giesinger et al. (2011) established CAT versions of the QLQ-C30 instruments. A preliminary list of 588 items was generated, based on a literature review. Then, after thoroughly selecting items, a pool of 44 items was attained, which was further revised via 52 patient semi-structured interviews [Giesinger 2011].

CAT procedure

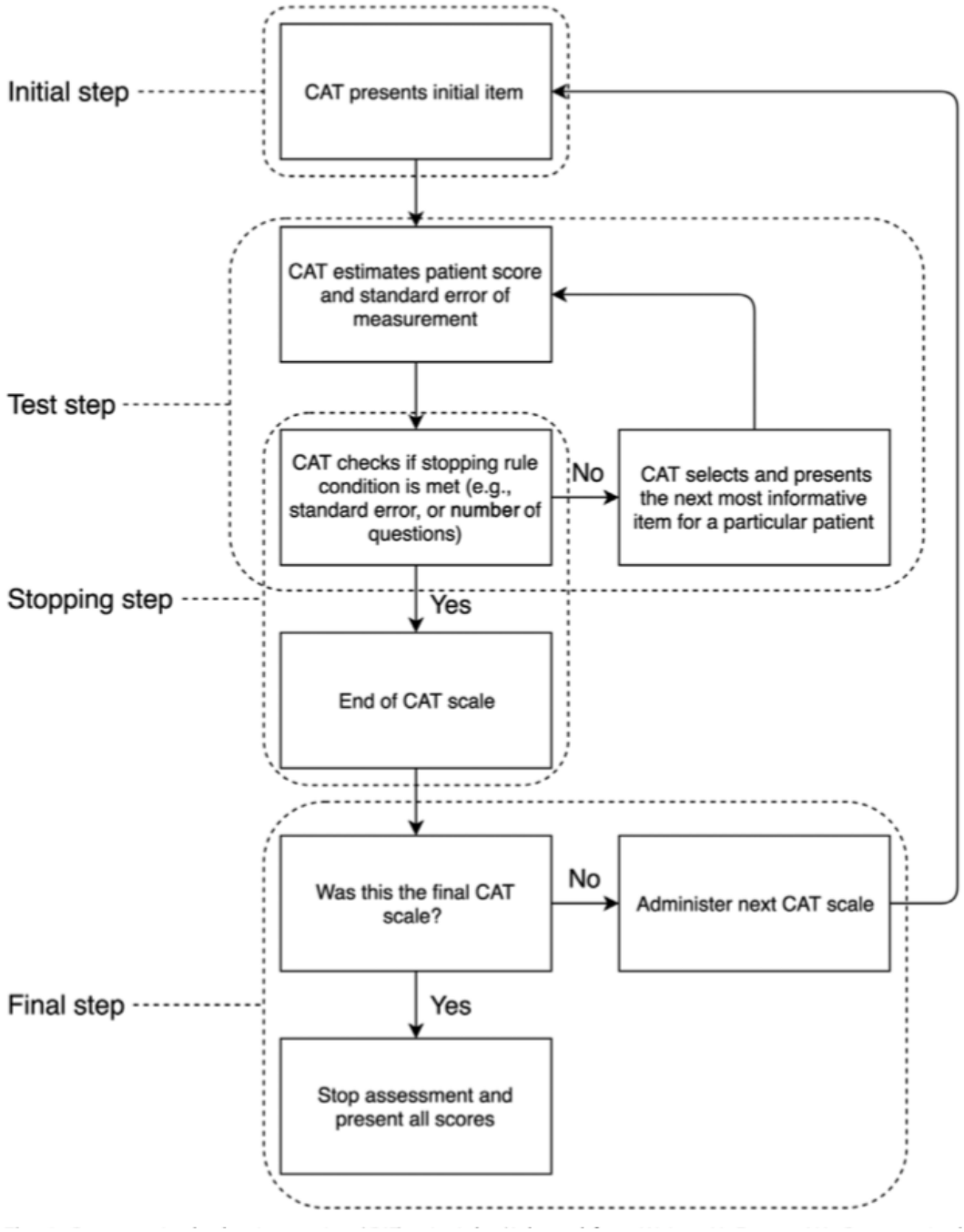


Figure 1. Computerized adaptive testing (CAT) principle (Adapted from Wainer et al., 2000, and Geerards et al., 2019).

Figure 1 above shows the CAT principle. As the patient starts the test, an initial item with medium difficulty is generally administered. Based on his/her answer, the initial estimate of the ability score is made. Then, the second item is selected, with difficulty based on the present estimation of the ability score. By responding to second item, it is possible to recompute the ability score with higher precision. By means of IRT, standard errors (SE) and confidence intervals (CI) are calculated. In case the SE is not deemed enough small, a further item is chosen from the item bank. This procedure goes forward, until adequate values are attained both for SE and CI. The stopping rule takes account of either the precision or the need to keep the test short. By simulations, it is possible to assess the effect of using different stopping rules. When simulations are generated using patient data are named real simulations [McBride 1997].

Evidence shows that CAT has been effectively used in education and psychology fields [Ware 2003; Wainer 2000]. Since nineties, many CATs were devised in health care setting, such as those for headache [Bjorner 2003; Ware 2003], rheumatoid arthritis [Martin 2007], osteoarthritis [Kosinski 2006], back pain [Kopec 2008], physical therapy [Jette 2007; Haley 2008], anxiety [Walter 2008], cancer [Petersen 2006; Petersen 2020], multiple sclerosis [Michel 2016], psychiatry [Michel 2018], and pediatrics [Allen 2008].

Moreover, the National Institute of Health promoted the Patient-Reported Outcomes Measurement Information System (PROMIS) developed a set of CATs for clinical research community [Reeve 2007]. So far, 3-7 item/4-10 item PROMIS instruments for fatigue, physical functioning, pain, depression, anxiety, and social function are available (www.nihpromis.org).

An example of the validation of a generic HRQOL CAT instrument comes from Rebollo et al. (2010), who showed that the HRQOL CAT-Health was valid and efficient in primary care patients.

An example of a recently published CAT in the cancer-related fatigue, comes from the EORTC QOL group who found that the EORTC CAT Core appraises the same dimensions as the original questionnaire without any floor or ceiling effects. In the EORTC CAT Core version, smaller samples are used without loss of power in comparison to the original questionnaire [Petersen 2020].

Another international project which needs to be mentioned is the Neuro-QOL, a project funded by the National Institutes of Health and National Institute of Neurological Disorders and Stroke. This project aimed to devise an HRQOL instrument that can be applied among different chronic neurologic diseases. As well as in the above-mentioned projects, an item bank was devised and additional disease-specific scales are developed, and made available. These tools are being administered using CAT [Miller 2016, Healy 2019].

In general, CAT presents several advantages, as follows: the item number is reduced, the item selection is individually tailored, thus omitting non pertinent items; floor and ceiling effects can be minimized; the exact degree of precision can be specified a priori, and test continues until that precision is attained; subjects producing incoherent responses can be identified; items and sub-groups of subjects reporting differential item functioning are promptly detected; physical problems of answer sheets are solved; scores are immediately available, allowing to provide immediate feedback to the patient, the clinician, and the researcher [Wainer 2000]. This has tremendous implications for using tests in clinical practice and research.

On the other hand, CAT presents also disadvantages, as follows: items pools should be devised and tested; for the Item calibration phase, large samples are required; these item pools need to be expanded and refined even once the CAT has been released; implementation in clinical practice is still difficult [Wainer 2000].

Multidimensional computerized adaptive testing

As described above, HRQOL instruments are generally multidimensional, and include multi-items scales. In multidimensional CAT (MCAT), as these scales are correlated, computations are complex, but the aim would be the same as for unidimensional CAT approach. In the MCAT an item could provide information regarding one or more latent variables. Thus, items are chosen to maximize information across abilities over all the dimensions [Wainer 2000]. MCAT may use these associations to improve measurement efficiency. In fact, Segall (1996) found that MCAT was more efficient than unidimensional CAT by reducing test length and increasing precision [Segall 1996].

A few other examples of use of MCAT in HRQOL are reported below.

Petersen et al. (2006) investigated multidimensional CAT of three EORTC QLQ-C30 scales (I.e. emotional functioning, physical functioning, and fatigue scales). Results indicated that multidimensional CAT improved instrument efficiency and precision.

Michel et al. (2016) devised the MusiQOL-MCAT, a multidimensional CAT version of the MusiQoL, based on a QOL questionnaire for people with multiple sclerosis. After MCAT simulations, a 16- item version was specified, as precision and accuracy were acceptable. Also, the external validity was adequate.

Further, Michel et al. (2018) developed the SQOL-MCAT, a QOL multidimensional CAT, based on the original (fixed-length) instrument for schizophrenia patients. Findings showed that **t**he MCAT-SQOL is 39% shortened in comparison to the original instrument, with acceptable precision and accuracy. Validity was established by means of correlations of the SQOL-MCAT scores and symptoms scores.

2. Multiple sclerosis and the (health-related) quality of life

Preface

After a short introduction on multiple sclerosis (MS), one of the most disabling chronic disorders of the central nervous system, an overview of the most-used MS HRQOL instruments will be summarized.

MS presents several symptoms affecting patient and caregiver lives, which can change over the course of the disease. Impact of such symptoms on HRQOL will be briefly described using pertinent examples from the MS literature. Then, use of HRQOL measures in MS clinical practice will be outlined with a list of some barriers and advantages which should be considered by investigators.

MULTIPLE SCLEROSIS

Multiple sclerosis (MS) is a degenerative and inflammatory chronic disorder of the central nervous system affecting both myelin and axons. Its onset is generally in early adulthood, typically in the third decade [Reich 2018]. This disorder is a multifactorial, varied, and immune-mediated disease affected by both genetic and environmental factors. Its prevalence varies across areas with the highest figures of 120 to 180 per 100,000 people described in Northern Europe, North America, and Australia [GBD 2016; Atlas of MS 2020]. About 80% of patients receive a diagnosis of relapsing-remitting MS. Relapses can lead to various clinical manifestations, with a range of characteristics, such as sensory disturbances or more severe symptoms. Over years, nevertheless, neurodegenerative component is likely to be more prominent than inflammation in causing disability progression. Symptoms most commonly reported are spasticity, fatigue, ataxia, and bladder dysfunction, followed by pain, depression, and loss of cognitive function [Gold 2003]. About

50% of people with MS eventually develop a secondary progressive form 15 years after diagnosis [Scalfari 2014]. About 10-15% of people with MS have a primary progressive form. Uncertainty is a steady feature of MS [Heesen 2011; NCC-CC 2004], as cause and mechanisms of the disease are largely inexplicable [Compston 2006, 2008]. Individual course and prognosis are varying and difficult to foresee, with about one-third of patients living with a 'benign' course [Degenhardt 2009; Ramsaransing 2006]. Prognostic information is a complex topic, and the information needs of people with MS with regard to this are rarely met [Dennison 2016; Dennison 2018]. An increasing number of disease-modifying drugs (DMDs) were licensed for clinically isolated syndrome, relapsing, secondary progressive, and recently also for primary progressive MS with the aim to reduce relapses and slow disease progression [Montalban 2018; Tramacere 2015]. To date, treatments are just partly effective, and long-lasting effects are still unknown [Ebers 2010; Freedman 2008; Shirani 2012]. Further, DMDs are costly and adverse effects are reported, causing low adherence [Bruce 2010; Tramacere 2015].

Considering the recently licensed and further upcoming DMDs, it is vital to provide people with MS with unbiased, up-to-date information to facilitate informed choice and shared decision making [Heesen 2011]. Besides DMD choices, there a number of other decisions people with relapsing MS should make as for example, relapse therapy, or motherhood choice [Prunty 2008; Köpke 2009].

HRQOL instruments in multiple sclerosis

Over the past 30 years the interest to assess outcomes in MS has increased considerably.

Among the standardized instruments which have been devised, the most- and widely-used by health professionals and researchers still remains the Expanded Disability Status Scale (EDSS) [Kurtzke 1983], that includes an activity limitations/impairment scale, supplemented with ambulation/mobility status.

Recently, attention has been given to assess MS outcomes using the patient perspective [Rothwell 1997]. Since 1992, publications on HRQOL and MS-specific instruments increased markedly [Solari 2005]. A number of reviews [Mitchell 2005; Solari 2005] on HRQOL in MS showed that there are some generic instruments which were used in MS [Rothwell 1997; Aronson 1997; Brunet 1996; Burden of illness 1998; Pfennings 1997; Lintern 2001], as well as disease-specific instruments which have been developed and validated [Vickrey 1995; Cella 1996; Fischer 1999; Rotstein 2000; Gold 2001; Hobart 2001; Ford 2001; Solari 1999; Vernay 2000; Acquadro 2003; Yamamoto 2004; Mendes 2004].

The MS-specific HRQOL instruments currently available are the following: the MS quality-of-life 54 (MSQOL-54) [Vickrey 1995]; the Functional Assessment of multiple sclerosis (FAMS) [Cella 1996]; the MS Quality of Life (MSQLI) [Fischer 1999]; the Hamburg Quality of Life Questionnaire in MS (HAQUAMS) [Gold 2001]; the RAYS [Rotstein 2000]; the MS Impact Scale-29 (MSIS-29) [Hobart 2001]; the MS International Quality of Life (MUSIQOL) [Simeoni 2008]; the Leeds MS Quality of Life (LMSQOL) [Ford 2001]. Of those, three include a generic core module (SF-36 [Vickrey 1005; Fischer 1999] or FACT-G [Cella 1996]) supplemented with an MS-specific module. MS patients were involved in the developmental phase [Solari 2005].

These questionnaires are available only in their original versions, except the MSQOL-54, that was translated into numerous languages [Vickrey 1995; Solari 1999; Vernay 2000; Acquadro 2003; Yamamoto 2004; Füvesi 2008], as well as the MUSIQOL, and the FAMS, that is also available in Portuguese [Mendes 2004]. Responsiveness was investigated in almost all questionnaires [Solari 2005].

The three most-widely used questionnaires are briefly described below.

The MSIS-29 includes 29 items distributed into two scales: psychological and physical, with 9 and 20 items, respectively. Each item is rated using a 5-point Likert scale, from independence to greater

compromise. The psychological scale score ranges from 5 (best) to 45 (worst), and the physical scale score ranges from 20 (best functioning) to 100 (worst functioning) [Hobart 2001].

The FAMS consists of 59 items including the FACT-G, a generic core measure, and additional MS specific items. The FAMS is divided into 6 scales: symptoms, mobility, general contentment, thinking/fatigue, emotional well-being, and family well-being. Scores range from 0 (worst HRQOL) to 176 (best HRQOL) [Cella 1996].

The MSQOL-54 includes the SF-36, plus 18 MS specific items. The whole MSQOL-54 consists of 52 items divided into 12 dimensions, plus two single items (sexual satisfaction and change in health). Just like the SF-36, physical and mental health composite scores are determined [Vickrey 1995].

Recently, in their review, Khurana et al. (2017) identified the MS-specific PROs and assessed the developmental phase, reliability and validity of these instruments, by means of Evaluating the Measurement of Patient-Reported Outcomes tool. Results showed that, among the PROs most frequently used in MS clinical trials, the MSIS-29 reported the best overall mean score, followed by the LMSQOL. Further, content validity of PROs in MS research is generally lacking.

Impact of MS symptoms on HRQOL

Evidence shows that MS has a substantial influence on HRQOL for MS patients at all stages of the disease. There are many factors, such as mood, coping, self-efficacy, and perceived support, which affect more the HRQOL of people with MS than the physiological variables, such as extent of MRI lesions or weakness [Mitchell 2005]. Additionally, fatigue and cognitive impairment are relevant predictors, even in people at earlier disease stage [Miller 2010].

Living with MS also affects patient physical and mental health, as well the health status of caregivers [Rivera-Navarro 2003; Mitchell 2005]. Depression symptom and cognitive compromise manifest

also in early disease phase, and affect negatively cognitive performance, mainly processing speed [Landrø 2004].

By using general- and disease-specific HRQOL questionnaires, Nortvedt et al. (2003) reported that HRQOL in MS is associated with disability, severe disease course, mental health problem, bladder and sexual problems, fatigue, and having a family member affected by MS. Further, in their review, Benito-Leon et al. (2003) found that HRQOL measures are strongly correlated with patient adjustment to MS, and disability. A number of studies reported that symptoms such as cognitive dysfunction, pain, bladder and sexual problems were all associated with lower HRQOL in MS patients [Morales-Gonzalez 2004; Sprangers 2000; Hakim 2000].

Further, Rothwell et al. (2007) pointed out that MS patients and their neurologists differed with regard to significance related to compromise in HRQOL dimensions.

Use of MS HRQOL measures in clinical practice

HRQOL data allow to make an overall evaluation of the patient's health status, which can be used as a base to tailor (pharmacological) interventions, and evaluate their effectiveness, either in the clinical trial setting or in routine care.

There are several advantages in including the HRQOL instruments in clinical practice, as follows: to detect disease features which can be usually overlooked; support health professionals recognize patient preferences, recommend or update treatment, foster communication between patient and physician, and encourage shared-decision making [Solari 2005; Valderas 2008]. Further, collecting clinical HRQOL trial data may help to obtain evidence which health professionals can fruitfully debate with their patients [Solari 2005].

On the other hand, some barriers have been acknowledged to the use of HRQL/PRO instruments in clinical practice, as follows: physician opinion with regard to these instruments; lack of theoretical

clearness in relation to the meaning of the instruments; and the real-world issues of data collection, scoring, and review [Mitchell 2005]. Furthermore, standard questionnaires can be unsuitable in a condition that is characterized by highly varied clinical manifestations, and compromise [Solari 2005]. Moreover, there is no gold standard which can serve to conduct assessment or report outcomes. Even the EMA in 2015 did report (EMA 2015): '*Few data are available on validation of specific instruments for QOL in patients suffering from MS. If evaluation of QOL in MS is considered, reliable and validated scales should be used. […] The development of patient reported outcomes is encouraged. Several patient reported outcomes are under evaluation. Their use and validity in multiple sclerosis should be justified in the study protocols. So far limited data are available. Hence specific recommendations on specific scales cannot be made'.*

3. Development of a Multidimensional Computerized Adaptive short form of the Multiple Sclerosis Quality of Life-54 (MSQOL-54-MCAT): an international collaborative project

Preface

Based on the theoretical background and evidence from the generic and MS-specific literature presented in the two chapters above, this chapter will report detailed results of an international collaborative project which aimed to develop a Multidimensional Computerized Adaptive Testing (MCAT) version of the MSQOL-54 inventory.

The project

The present retrospective, cross-sectional project is part of an international collaborative initiative of Italian and Australian researchers who set up and share a unique international (and multiple-language) MSQOL-54 database, in order to develop a computerized adaptive version of the MSQOL-54 using data collected in different countries. This project is independent, and relies on a fruitful collaboration between investigators with different expertise coming from different countries. If our findings will be positive, the next step could be to add/integrate items to the original item pool (not part of the present thesis), and investigate the possibility of producing a unidimensional outcome measure derived from the MSQOL-54 (Leader: Dr. Jelinek, University of Melbourne, not part of the present thesis).

INSTRUMENT

As briefly described in the chapter 2 above, the MSQOL-54 inventory was designed with the goal of comprehensively assessing the HRQOL of patients with MS. Compared to other instruments, its main strength is that it combines a generic- and a disease-targeted approach. In fact, it is a multidimensional, MS-specific HRQOL instrument, based on the generic SF-36 [Ware 1993] supplemented with 18 MS-specific items [Vickrey 1995]. It consists of 52 items combined in 12 subscales, and two single items (Table 1). Two composite scores (Mental Health Composite, MHC, and Physical Health Composite, PHC) are determined by aggregating scores of the pertinent subscales [Vickrey 1995]. Psychometric properties like construct and content reliability, discrimination [Solari 1999; Idiman 2006; El Alaoui 2012], and responsiveness [Giordano 2009] have been rigorously documented. It was developed in US English, and clinically validated in various languages [Solari 1999; Füvesi 2008; Idiman 2006; El Alaoui 2012; Acquadro 2003; Yamamoto 2004; Pekmezovic 2007; Füvesi 2008], including Italian [Solari 1999].

Table 1. MSQOL-54 items and subscales.

Items	Scales	Summary measures
3. Vigorous activities 4. Moderate activities 5. Lift, carry groceries 6. Climb several flights 7. Climb one flight 8. Bend, kneel 9. Wilk mile 10. Walk several blocks 11. Walk one block 12. Bath, Dress	Physical health	**PHYSICAL HEALTH COMPOSITE**
1. EVGFP rating 34. Sick easier 35. As healthy 36. Health to get worse 37. Health excellent	Health Perceptions	
23. Pep/Life 27. Energy 29. Worn out 31. Tired 32. Rested on walking in the morning	Energy	
13. Cut down time 14. Accomplished less 15. Limited in kind 16. Had difficulty	Role limitations due to physical problems	
21. Pain magnitude 22. Pain interfere with work 52. Pain interfere with enjoyment	Pain	
46. Lack of sexual interest 47. Erection/Lubrication 48. Orgasm 49. Satisfy sexual partner	Sexual function	
20. Social extent, physical health 33. Social time 51. Social extent, bowel or bladder	Social function	
38. Discouraged 39. Frustrated 40. Worried for life 41. Weighed down	Health distress	
38. Discouraged 39. Frustrated 40. Worried for life 41. Weighed down	Health distress	**MENTAL HEALTH COMPOSITE**
53. 0-10 NRS (worst possible-best possible) rating 54. TUMMMPO rating	Overall quality of life	
24. Nervous person 25. Down in dumps 26. Peaceful 28. Blue/Sad 30. Happy	Emotional well-being	

17. Cut down time 18. Accomplished less 19. Not careful	Role limitations due to emotional problems	
42, Concentration and thinking 43. Sustained attention 44. Memory 45. Others note troubles with memory or concentration	Cognitive function	
2. MSASM rating	Change in health	-
50. Satisfied with sexual function	Satisfaction with sexual function	-

EVGFP, Excellent, Very good, Good, Fair, Poor. NRS, Numeric Rating Scale. MSASM, Much better now than one year ago, Somewhat better now than one year ago, About the same, Somewhat worse than one year ago, Much worse now than one year ago. TUMMMPO, Terrible, Unhappy, Mostly dissatisfied, Mixed – about equally satisfied and dissatisfied, Pleased, Delighted.

PARTICIPANTS

We collected different datasets with the English and Italian language versions of MSQOL-54 within ongoing or completed research projects carried out in Australia and Italy. These datasets constitute the data which were included in this thesis.

English version data - We obtained the English version data of the MSQOL-54 from the '*HOLISM study*': Australian investigators started and coordinated this observational international study, whose methods and findings have been described in details elsewhere [Hadgkiss 2013; Jelinek 2016]). Briefly, people with MS coming from Australasia, Europe, North America, and other countries have been enrolled using web-based platforms, including social media, websites and forums involving people with MS. The HOLISM study presents an overview of current lifestyle, habits, and risk-modifying behaviors of a large international sample of people with MS, as well as an ongoing platform to assess longitudinally the association between these variables and disease and symptom progression. For the thesis, we included data from English-speaking countries only: 840 (41%) from North-America, 797 (39%) from Australasia, and 427 (20%) from UK & Ireland.

Italian version data – We obtained the Italian version data from the following sources:

- The '*Care system project*' [Bassi 2014; Bassi 2016], an observational study about patient's perceived levels of ill-being and well-being (Italian Multiple Sclerosis Foundation, FISM grant number 2011/R/5), and a larger study (FISM grant number 2014/R/4). For the thesis, we included data of 662 people with MS from 8 MS centers.
- The study '*An abbreviated computerized version of the MSQOL-54: Development and preliminary validation using Confirmatory Factor Analysis and Item Response Theory*' (FISM grant number 2013/R/20) [Rosato 2016; Rosato 2018; Massacesi 2014; Solari 2004], in which an abbreviated version of the MSQOL-54 was devised. For the thesis, we included data from 564 people with MS (from 5 MS centers) who participated in the retrospective phase of the study [Rosato 2016].
- Other research projects carried out in 5 Italian MS centers. For the thesis, we included 379 people with MS.

Ethics committees approvals - All the projects mentioned above have been approved by local ethics committees (St Vincent's Hospital Melbourne Human Research Ethics Committee [LRR 055/12]; Università di Torino; Università di Milano; San Raffaele Hospital, Milano; University Polyclinic Hospital G. Rodolico, Catania; University of Florence; S. Anna Hospital, Como; Hospital of Vaio-Fidenza, Fidenza; University 'G. d'Annunzio', Chieti; University of Bari; San Camillo-Forlanini Hospital, Rome; University Hospital 'San Luigi Gonzaga', Orbassano; Fondazione IRCCS Istituto Neurologico 'C. Besta', Milano; IRCCS S. Lucia Foundation, Rome). Patients gave written or online informed consent to be included in the original projects. Additional consent was not required for this secondary analysis, for which patients' privacy and anonymity were guaranteed.

AIMS AND ACTIONS

The main aim of the present project was to develop a MCAT version of the MSQOL-54. In doing so, the first action was to assess whether was possible to merge Italian and English language versions MSQOL-54 data presented above. The second action was to apply the bifactor model to the MSQOL-54 items in order to verify whether a total HRQOL score could be calculated. The third action was to apply multidimensional CAT to the MSQOL-54, and assess its performance, in comparison to the fixed-length questionnaire.

ACTION 1 – Assessment of the measurement invariance of MSQOL-54 across Italian and English versions

The methods and the results of this Action are fully reported in a recently published paper [Giordano 2020].

Database set up

Before starting to merge the data coming from different language versions of the MSQOL-54, it was necessary to perform extensive data quality checks. These checks consisted of a detailed search for possible multiple imputations coming from the same subject. We searched for the records which had the same date of birth and sex, both within and across datasets, and we removed duplicates.

The database records were eligible if the following variables were available: MS diagnosed according to McDonald's [Polman 2005]/McDonald's revised criteria [Polman 2011] (Italian version), or if the diagnosis was posed by a physician (English version), and if patient age was ≥18 years, gender, disease duration, EDSS [Kurtzke 1983] (Italian version only), and Patient Determined Disease Steps scores (PDDS [Hohol 1995]). We included database records when more than 67% of the MSQOL-54 items were completed.

Before analyzing the data provided by the two datasets together, we wondered whether, from a psychometric point of view, the instrument works in the same way in both the Italian and English language administrations. With the objective to check whether was possible to pool data coming from the Italian and English language versions of the MSQOL-54, we assessed the measurement invariance of the two language versions of the MSQOL-54.

A note on measurement invariance

Measurement invariance is a relevant statistical property of an instrument attesting that the same latent construct is assessed across time/groups [Putnick 2016]. The failure of invariance to hold is, in most situations, evidence that the manifest variables (e.g. ability to climb stairs, walk a few miles, presence of pain) fail to measure the same latent attributes (e.g. physical functioning) in the same way in different situations or subgroups. Unless measurement invariance has been demonstrated, it is not possible to perform meaningful cross-group comparisons. Pooling data across samples collected in different countries with different languages may be problematic, as specific cultural beliefs and expectations may affect the interpretation of items; differences in observed scores may thus not reflect actual differences in latent variables. Lack of measurement invariance across versions - as well as across cultural contexts - can be due to poor translation or because items are not applicable across cultures, elicit further concepts or present ambiguous nuances [Boer 2018].

In the MS field in general, few studies have assessed measurement invariance of instruments [Motl 2010, 2011, 2012; Cox 2014; Chung 2015, Chung 2016]. Among these, the majority have evaluated measurement invariance across groups, with small sample sizes, and analyzed data using multi-group confirmatory factor analysis [Motl 2011, Motl 2012; Cox 2014; Chung 2015, Chung 2016].

To the best of our knowledge no study has evaluated measurement invariance of MSQOL-54 across language versions. According to recent studies which found evidence of partial invariance in HRQOL

instruments [Byrne 1989; Santos 2017; Geyh 2010], and that Italian and English are western languages, we expected that full or at least partial invariance would hold across the two language versions.

METHODS

Analysis

Further, CFA with full information maximum likelihood estimation (FIML) with robust standard errors was used to separately assess whether the data from the two language versions fitted the original MSQOL-54 12-factor model [Vickrey 1995], and then to assess measurement invariance across the two language versions. FIML is one of the approaches to dealing with missing data. Under FIML, instead of imputing the values of missing data and then determining the value for the unknown parameters, all the available data from complete and incomplete records are used to produce parameter estimates. Assuming the missing at random (MAR) condition, FIML tends to be approximately unbiased in large samples and it is also highly efficient [Schafer 2002].

The factor structure of MSQOL54 that we tested in the present work is the one presented by Vickrey in 1995, consisting of 52 items combined in 12 subscales (see Table 1 above).

In CFA model the parameters can be freely estimated, fixed, or constrained. A free parameter is unknown, and the researcher allows the algorithm to find its optimal value that, together with other model estimates, minimizes the differences between the observed and predicted variance–covariance matrices (e.g., in a one-factor CFA model, to obtain the set of factor loadings that best reproduces the observed correlations among four input indicators).

The most basic requirement of questionnaire factorial invariance is that in a set of populations (e.g., Italian speakers and English speakers), there exists an invariant factor loading matrix. If the factor

loadings are equivalent (invariant), the magnitude of the relationships between the items and the underlying construct (e.g., physical QOL) are the same in different subgroups.

A fixed parameter is specified by the researcher to be a specific value. Like a free parameter, a constrained parameter is unknown. However, the parameter is not free to be any value; rather, the specification places restrictions on the values it may assume. The most common forms of constrained parameters are equality constraints, in which unstandardized parameters are restricted to be equal in value. In multi-group CFA, when testing for invariance between groups, we estimate the CFA parameters freely on one group and constrain the parameters in the other group to be equal. Parameters estimated in CFA models are factor loadings, unique variances, and factor variances. Factor loadings are the regression slopes and indicate the impact that the latent variable has on each indicator. The unique variance is the variance of the manifest or indicator variable that is not represented by the latent variables, and usually is referred to as error variance and indicator unreliability. When the CFA solution consists of two or more factors, a factor covariance is also specified to estimate the relationship between the latent dimensions, although factors covariance may be fixed to zero. The CFA model may also include a mean structure analysis and try to reproduce the observed sample averages of the manifest variables. As a result, such CFA models also include parameter estimates of indicator intercepts (expected value of the indicator when the factor is zero) and averages of latent variables, which are often used in multi-group CFA to test whether distinct groups differ in their relative position on the latent dimensions.

Three increasingly constrained levels of measurement invariance (i.e. configural, metric, and scalar) were assessed by constraining specific parameters in each instance [Millsap 2004; Giordano 2020]. In particular, configural invariance which tests whether the same pattern of loadings exists across the groups under investigation (i.e., Italian and English language versions), requiring that the same items have non-zero loadings on the same factors. If the pattern loadings are not equivalent across

groups, then configural invariance fails; the indicators variables, i.e., physical functioning items are not 'reflecting' the latent factors in the same way across groups, and meaningful group comparisons based on manifest variates cannot be made. In this case, the same factors cannot be assumed to underlie the manifest items.

The metric invariance assumes that factor loadings are equal across groups. If metric invariance fails, the items load on the same latent factors but with "different" degree on the two language samples. Scalar invariance produces a model in which, besides the factorial weights, items' intercepts (i.e., differences in item intercepts) and residuals (i.e., differences in the amount of variation left unexplained in each item by its respective latent) are also held equal across groups. A model is considered suitable if the covariance structure implied by the model is similar to the covariance structure of the sample data.

Several indices measure the goodness of fit of the model to the data. Such fit indices can be classified into absolute and incremental fit indexes. In the present analysis, we used two absolute fit indices (i.e. root mean square error of approximation, RMSEA; and standardized root mean square residual, SRMR); and one incremental fit index (i.e. Comparative Fit Index, CFI). Their detailed formulas are reported below:

Root mean square error of approximation (RMSEA)

$$RMSEA = \sqrt{\frac{\chi_t^2 - df_t}{df_t\ (N-1)}}$$

Here, χ^2 corresponds to the chi-square for the tested model, df to the degrees of freedom, and N to the sample size. RMSEA is the discrepancy between the observed covariance matrix and covariance matrix implied by the model, per degree of freedom [Steiger 1980; Browne 1993].

Standardized root mean square residual (SRMR)

$$SRMR = \sqrt{\frac{2\sum\sum\left[\left(s_{ij} - \sigma_{ij}\right)/(s_{ii}s_{jj})\right]^2}{p\,(p+1)}}$$

Here, s_{ij} corresponds to the observed covariance, σ_{ij} to the covariance implied by the model, s_{ii} and s_{jj} to the observed SDs, and *p* to the number of observed variables. SRMR is calculated as the average of the standardized residuals between the observed and covariance matrices implied by the model [Bentler 1995].

The Comparative Fit Index (CFI)

$$CFI = 1 - \left\{\frac{\chi_t^2 - df_t}{\chi_n^2 - df_n}\right\}$$

Here, χ_t^2 corresponds to the chi-square for the tested model, χ_n^2 to the chi-square for the null model, and df_t and df_n to the degrees of freedom for the tested and null models, respectively. CFI assesses the extent to which the tested model is superior to an alternative model in reproducing the observed covariance matrix [Bentler 1990; McDonald 1990].

To consider the model fit acceptable, we used the cut-off criteria reported by Hu et al. (1995) and Hu et al. (1999), as follows: RMSEA <0.08; CFI >0.90; and SRMR <0.08.

As measurement invariance assessment consists of increasingly constrained levels, we compared the fit of the nested models (metric, scalar, and configural invariance), by calculating the difference between fit statistics for such models (e.g., Δχ2, ΔCFI, ΔRMSEA, ΔSRMR).

According to Chen [Chen 2007], a worsening of CFI that exceeds the threshold of 0.010, supplemented by a change of ≥ 0.015 in RMSEA or a change of ≥ 0.030 in SRMR was considered as

indication of absence of metric invariance; when testing scalar invariance, the cut-off values for CFI and RMSEA were the same as for metric invariance, while it was 0.010 for SRMR [Hays 2005].

In addition, as there were significant differences between socio-demographic and clinical characteristics between the two samples, and within the English-language version sample MS patients came from different geographic areas (i.e. North-America, Australasia, and UK & Ireland), we cannot exclude that the above mentioned differences could lead to non-invariant parameters. Therefore, two sensitivity analyses were performed to account for possible sample selection biases, by assessing measurement invariance considering the following:

a) English-speaking geographic areas (Australasia/North-America/UK & Ireland); here, we assessed whether responses underlying the latent construct of people with MS coming from English-speaking countries would be different across geographical areas.
b) To assess whether the latent constructs would be the same across two sub-samples of people with MS (N=985 each) matched for gender, age (18-30 years, 31-40, 41-50, 51-60, 61+), level of disability, and disease duration (0-11 years, 12-23, 24+) by using 1:1 coarsened exact matching [Iacus 2009].

All analyses were performed with Stata Statistical Software, release 12.0 (Stata Corp LP, College Station, USA), and Mplus software 7.0 [Muthén 2012].

RESULTS

Descriptive analysis

The original dataset (including the two language versions) comprised 3877 people with MS. Of those, 37 were excluded as they were duplicates, 96 because they did not complete any MSQOL-54 item, and 75 because they completed less than 67% of the items. Out of the 3669 MS patients who

were included, 1605 (44%) were Italian (mean age 41 years, 62% women, 69% with a mild disability level) and 2064 (56%) were English-speaking (840 [41%] from North-America, 797 [39%] from Australasia, 427 [20%] from UK and Ireland, with mean age 46 years, 83% women, 54% with a mild disability level). Compared to Italians, English-speaking participants were older, had a higher percentage of women, and had longer disease duration (p< 0.001) (Table 2; Appendix).

Table 2. Characteristics of the entire dataset (N=3669 patients) by MSQOL-54 language version.

	English-speaking (N=2064)	**Italian (N=1605)**	**P value**
Women (%)[1]	1704 (83)	996 (62)	<0.001
Mean age in years, SD (range)[2]	46.1, 10.5 (18–87)	40.9, 10.8 (18–79)	<0.001
Mean years from MS diagnosis, SD (range)[3]	9.0, 7.3 (1–42)	4.9, 7.8 (0-48)	<0.001
Median EDSS score (range)[4]	-	2.5 (0–9.5)	-
Patient Determined Disease Steps (%)[5]			
Mild disability	1110 (54)	1097 (69)	
Moderate disability	722 (35)	308 (19)	
Severe disability	219 (11)	194 (12)	<0.001
Mean MSQOL-54 PHC, SD (range)	57.7, 21.5 (3–100)	61.1, 20.2 (2–100)	<0.001
Mean MSQOL-54 MHC, SD (range)	66.6, 21.3 (1-100)	62.9, 20.7 (2–100)	<0.001

EDSS, Expanded Disability Status Scale; MSQOL-54, Multiple Sclerosis Quality of Life-54; PDDS, Patient Determined Disease Steps; PHC/MHC, Physical and Mental Health Composite; SD standard deviation.

1. Missing replies for sex: N=21 (English-speaking).
2. Missing replies for age: N=62 (English-speaking); N=53 (Italy)
3. Missing replies for disease duration: N=11 (English-speaking); N=227 (Italy)
4. Missing replies for EDSS: N=6 (Italy).
5. Missing replies for PDDS: N=13 (English-speaking); N=6 (Italy).

Measurement invariance

The 12-factor model of the MSQOL-54 was estimated separately in the two language versions, using the maximum likelihood estimation with robust standard errors, obtaining good fit indices for RMSEA and SRMR (Italian: RMSEA=0.050; SRMR=0.045; English: RMSEA=0.054; SRMR=0.047), and

an acceptable value for CFI (Italian: CFI=0.906; English: CFI=0.903). The model assessing the configural measurement invariance produced analogous results to those in the separate samples: good fit indices for RMSEA and SRMR and a less satisfactory, but still acceptable, value for CFI (Table 3). For the model in which loadings were constrained to be equal across groups, the fit indices were acceptable and the worsening with respect to the unconstrained model was negligible (ΔRMSEA<0.001; ΔCFI=-0.002, ΔSRMR=0.002), supporting the metric invariance of the instrument. Finally, when both loadings and intercepts were constrained to be equal across groups (scalar invariance), the model fitted the data well in terms of RMSEA and SRMR, and CFI was slightly under the cut-off of 0.90. Concerning the changes in fit indices as compared to the metric invariance model, the cut-off values were reached, except for ΔCFI (ΔRMSEA=0.003; ΔCFI=-0.013, ΔSRMR=0.003), supporting scalar invariance.

Table 3. Measurement invariance of the MSQOL-54.

	χ^2(df)	χ^2 p-value	RMSEA	CFI	SRMR	Δ RMSEA	Δ CFI	Δ SRMR
Italian (N=1605)	5987.5 (1208)	<0.0001	0.050	0.906	0.045	-	-	-
English-speaking (N= 2064)	8596.3 (1208)	<0.0001	0.054	0.903	0.047	-	-	-
Configural invariance	14508.0 (2416)	<0.0001	0.052	0.904	0.046	-	-	-
Metric invariance	14829.6 (2456)	<0.0001	0.052	0.902	0.048	0.000	-0.002	0.002
Scalar invariance	16551.8 (2496)	<0.0001	0.055	0.889	0.051	0.003	-0.013	0.003

CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

Sensitivity analysis

Measurement invariance was also assessed across English-speaking geographic areas (North-America/Australasia/UK & Ireland). Results supported configural, metric, and scalar invariance across the three subgroups, indicating that the loadings and intercepts of the MSQOL-54 items can be considered equal across the different English-speaking areas (Table 4).

Table 4. Measurement invariance of MSQOL-54 across geographic areas within English-speaking participants.

	χ^2(df)	χ^2 p-value	RMSEA	CFI	SRMR	Δ RMSEA	Δ CFI	Δ SRMR
Australasia (N= 797)	4067.4 (1208)	0.0000	0.054	0.897	0.051			
UK & Ireland (N= 427)	2932.3 (1208)	0.0000	0.058	0.894	0.054			
North-America (N=840)	4290.3 (1208)	0.0000	0.055	0.905	0.052			
Configural invariance	11316.5 (3624)	0.0000	0.056	0.900	0.052			
Metric invariance	11488.6 (3704)	0.0000	0.055	0.899	0.054	-0.001	-0.001	0.002
Scalar invariance	11724.0 (3784)	0.0000	0.055	0.897	0.054	0.000	-0.002	0.000

CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

Results from the matched-pairs subgroup analysis supported configural, metric, and scalar measurement invariance, indicating that the results of the main analysis reported in Table 2 were not biased by the demographic and clinical differences across the language version samples (Table 5).

Table 5. Measurement invariance of MSQOL-54 across two sub-samples matched for age, sex, level of disability and disease duration.

	χ^2(df)	χ^2 p-value	RMSEA	CFI	SRMR	Δ RMSEA	Δ CFI	Δ SRMR
Configural invariance	9373.6 (2416)	<0.0001	0.054	0.899	0.050			
Metric invariance	9566.1 (2456)	<0.0001	0.054	0.896	0.053	<0.001	-0.003	0.003
Scalar invariance	10445.8 (2496)	<0.0001	0.057	0.884	0.055	0.003	-0.012	0.002

CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

DISCUSSION

Measurement invariance is an important prerequisite for meaningful group comparisons. To the best of our knowledge, this is the first study investigating the measurement invariance of the MSQOL-54 across language versions. Our findings support measurement invariance of the English and Italian MSQOL-54, suggesting that the questionnaire has the same meaning across languages, and that individuals who have the same score on a MSQOL-54 domain would obtain the same value on the observed variable, irrespective of the language version.

In the sensitivity analysis we found that measurement invariance was further supported across English-speaking countries, which is important considering that the original US English version of the MSQOL-54 was used in all these countries. Further, measurement invariance was supported across subgroups matched for age, sex, level of disability, and disease duration.

Overall, these findings indicate that the MSQOL-54 can be used to assess HRQOL among both Italian- and English-speaking people with MS. They further demonstrate that it is possible to pool data or compare scores between these two language groups (and within English-speaking groups) and obtain meaningful interpretations. Any perceived similarities or differences in HRQOL levels

between Italian- and English-speaking people with MS would therefore indicate true similarities or differences. Notably, the (original) US English version (used with English-speaking participants from the 'HOLISM study') and the Italian version, which has been linguistically validated according to international guidelines, can be considered culturally equivalent. The UK English version of the questionnaire has not yet been validated. However, it is not always feasible to validate an instrument in each target language group, so its validity in our populations is encouraging and produces evidence to support using the MSQOL-54 in other English-speaking populations.

As far as the methods of analysis are concerned, we chose multi-group confirmatory factor analysis because it is one of the most powerful analytical approaches in cross-cultural research. Given the response structure of some MSQOL-54 items (i.e. 2/3/4/5/6 response options), an estimation method for ordered response categories (e.g. weighted least square mean and variance adjusted estimator [WLSMV] using the polychoric correlation) may have been more appropriate [Millsap 2004]. However, no statistical methods other than $\chi 2$ are currently available to assess the measurement invariance between nested models when WLSMV is employed. Criteria for changes in CFI and other goodness of fit (GFI) indices have not yet been set, and the few studies addressing this issue suggest users avoid interpreting the changes in GFI, especially for mis-specified models [Sass 2014]. Moreover, in the present study we used a large dataset and it is known that the $\chi 2$ test statistic is sensitive to sample size, such that it tends to yield significant results [Hays 2005].

This study has some limitations. First, differences must be acknowledged in the recruitment strategies adopted to gather Italian and English data. Particularly, Italian data stem from research projects where clinical information was provided by investigators. By contrast, English data were derived from an online survey requiring a high level of literacy of participants. Moreover, higher levels of physical disability may have prevented some people with MS from participating and completing the survey without support. Further, some people with MS were directly recruited

through a website and associated forums promoting lifestyle changes; this may have facilitated the participation of individuals with a specific interest in this topic. In spite of these differences, our results globally support the robustness of the questionnaire.

Second, in the two datasets, disability level was assessed using different scales, the EDSS in Italy and the PDSS in the English-speaking population. To overcome this issue, EDSS scores were transformed into PDDS levels [Hohol 1995; Marrie 2005; Marrie 2006], improving the completeness of the data collected.

Third, other potential variables (such as level of education, employment, and disease form) were not available in the two original datasets; we therefore could not take them into account in data analysis.

CONCLUSION

To conclude, results from this study further support the inclusion of the MSQOL-54 as a PRO in clinical practice and research involving both Italian- and English-speaking people with MS. Moreover, findings show that data gathered with these language versions can be suitable for group comparisons and can be pooled to obtain large international datasets needed to apply the multidimensional computerized adaptive testing to the MSQOL-54.

Future studies should be conducted to further assess measurement invariance across language version groups matching the samples by a broad set of individual and clinical variables, such as levels of education, employment, and disease forms. Taking into account those variables would increase confidence that comparisons across language versions are meaningful.

Finally, researchers have recently shown substantial interest in using electronic PROs to routinely monitor patients with long-term conditions. One step forward could be to assess measurement

invariance across the modes of MSQOL-54 administration (paper vs. electronic) in both Italian and English versions of the instrument.

ACTION 2 - Viability of a MSQOL-54 general health-related quality of life score using bifactor model

Preface

Results obtained in the Action 1 above, by assessing the measurement invariance across the Italian and English language versions of the MSQOL-54, show that these data can be pooled to obtain a unique, large, international dataset. The final goal of this thesis was to attempt to provide a MCAT version of the MSQOL-54 inventory. Most applications of CAT used unidimensional item response theory (IRT) models. These IRT models are appropriate for many psychological variables which account for individual differences along a single psychological dimension (e.g. mathematical ability, depression disorder, etc.). However, some psychological variables are multidimensional and, among these, HRQOL is definitely a multidimensional construct. When a psychological variable is multidimensional, there are two general approaches to modeling it with IRT. The first is to use the multidimensional IRT models [Bock & Aitkin 1981; Bock 1988; Reckase 1985; Reckase & McKinley, 1991], which estimate parameters for each item that describe the item's contribution in measuring the underlying latent variables. Once these item parameters are estimated and the item bank has been calibrated, CAT could proceed by applying multidimensional CAT algorithms along with the multidimensional item parameters [e.g., Segall 1966; Segall 2000; van der Linden 2000]. One possible drawback with this approach is that multidimensional IRT may not result in estimates of the trait (θ) that is assumed to underlie the "general" variable under study (i.e. HRQOL).

A plausible alternative factor structure is the “bifactor” model [Holzinger & Swineford 1937] that constrains each item to have a non-zero loading on the “general” dimension (e.g. HRQOL) and a secondary loading on no more than one of the domain contents factors (e.g. physical functioning). The bifactor structure is plausible in HRQOL measurement, where symptom items that are related

to a primary dimension of interest are often selected from underlying measurement sub-domains. In 2007, Gibbons et al. extended the bifactor model for analysis of graded response data. The advantage of the bifactor model is that it yields an overall or "general" measure that can be the focus of CAT, as well as measurement on the underlying sub-domains. However, the bifactor model is a confirmatory model that starts with the factor structure that is assumed to underlie the construct to be measured. The bifactor model can be estimated within the logic of IRT models by converting the parameter estimates of the bifactor model into the parameters of an IRT model with the goal of calibrating the item bank before implementing the CAT model. In the bifactor model, each item loads simultaneously on the general factor, and on one of the group factors. The results of the present Action 2 reported below are under review in a peer-reviewed journal.

Introduction

The idea of estimating the factorial structure of the MSQOL-54 by means of the bifactor model arose from observing some empirical results on the second-order multifactorial structure proposed by Vickrey et al. (1995) in the developmental work on the MSQOL-54. In that work, a quite high correlation (r=0.66) was reported between the two MSQOL-54 composite factors (i.e. mental health composite and physical health composite). Given this correlation, it could be hypothesized that a unique total score of HRQOL may be calculated, with the benefit to provide patients, clinicians and researchers with a single overall HRQOL assessment, to assess for example, treatment response or modify treatment plan. In this very context, applying a bifactor model to the MSQOL-54 items could be particularly useful, as it is intended to acknowledge multidimensionality and, at the same time, take account of a single general construct [Chen 2006], as the HRQOL related to MS is. Applying the bifactor model to the MSQOL-54 items may constitute an alternative to the more widely-used second-order models, or correlated-traits [Reise 2010]. By definition, the bifactor model is

employed so that items load on one group factor only, and the general and group factors are all uncorrelated to each other [Reise 2010].

Bifactor modeling is generally used to test multifaceted constructs [Chen 2006], and so far, has been used mainly in the area of intelligence research [Gustafsson 1993; Luo 1994], and in the study of personality [Bludworth 2010; Brouwer 2008]. However, this has rarely been applied in neurology and MS research, except for a few studies [Chilcot 2016; Chamot 2014; Mokkink 2011].

In the present Action 2, our primary aim was to apply the bifactor model to the MSQOL-54 items in order to verify whether a total HRQOL score could be calculated. Second, if the bifactor model fitted the data well, we aimed to evaluate the measurement invariance of MSQOL-54 items across age and gender.

METHODS

To perform the present secondary analysis, we used the data presented in the Action 1 above, and in a recent publication [Giordano 2020].

Statistical analysis

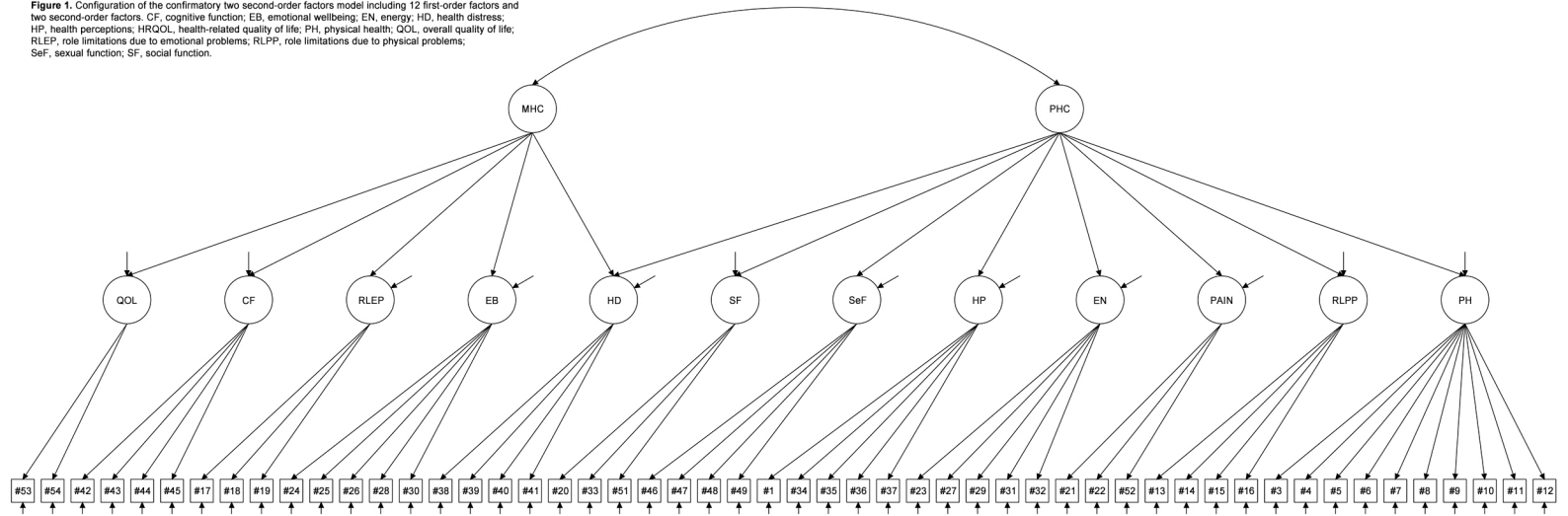
The goodness of fit of the original second-order factor model comprising two factors, the novel second-order factor model comprising one factor, and the bifactor model was tested using confirmatory factor analysis (CFA). As in the Action 1 above, full information maximum likelihood estimation with robust standard errors was used in order to account for the missing data and the violation of multivariate normality.

According to the original factor structure of the MSQOL-54, in the two second-order factor model, it was hypothesized that 52 items loaded in 12 first-order factors and two second-order factors, corresponding to the PHC and MHC [Vickrey 1995] (Figure 1). The remaining two items (i.e. item 2

'Compared to one year ago, how would you rate your health in general now?', and item 50 *'Overall, how satisfied were you with your sexual function during the past 4 weeks?'*) were not included in this model, as well in the other models, because they are single items.

In the single second-order factor model, the first-order factors were the same as in the original model, and one second-order factor was imposed, called 'HRQOL' (Figure 2). In the bifactor model, it was hypothesized that 50 items loaded onto the general HRQOL factor and on their specific group factors, whereas the two items forming the overall QOL subscale (items 53 and 54) were loaded only onto the general factor, because the bifactor model needs each group factor to be composed of at least three items to be identified (Figure 3).

Figure 1. Configuration of the confirmatory two second-order factors model including 12 first-order factors and two second-order factors. CF, cognitive function; EB, emotional wellbeing; EN, energy; HD, health distress; HP, health perceptions; HRQOL, health-related quality of life; PH, physical health; QOL, overall quality of life; RLEP, role limitations due to emotional problems; RLPP, role limitations due to physical problems; SeF, sexual function; SF, social function.



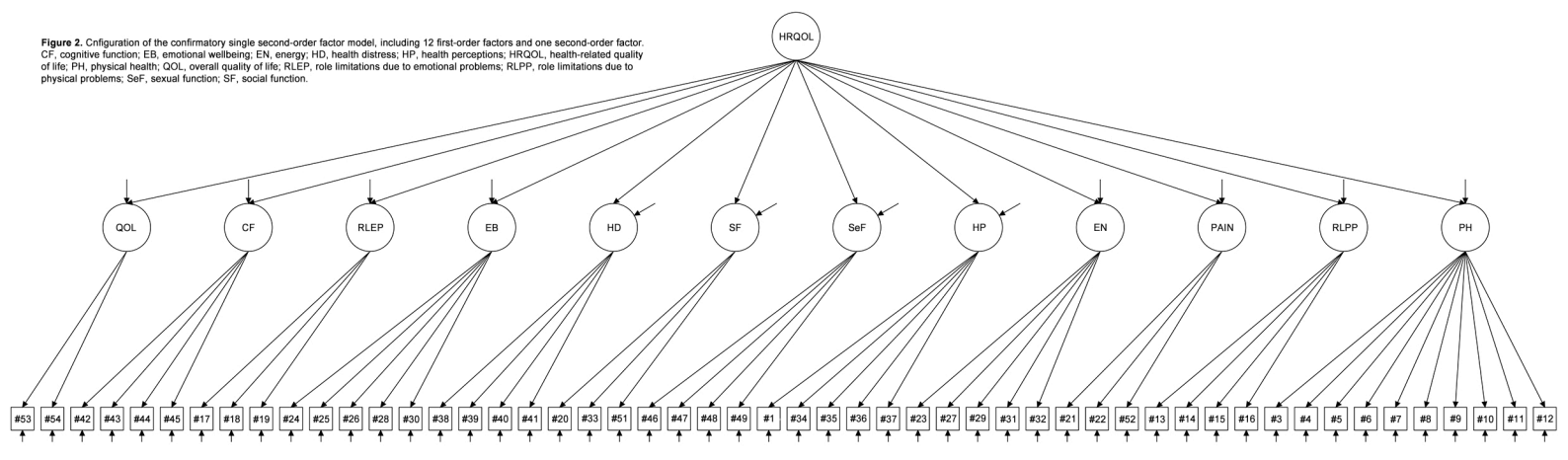


Figure 2. Cnfiguration of the confirmatory single second-order factor model, including 12 first-order factors and one second-order factor. CF, cognitive function; EB, emotional wellbeing; EN, energy; HD, health distress; HP, health perceptions; HRQOL, health-related quality of life; PH, physical health; QOL, overall quality of life; RLEP, role limitations due to emotional problems; RLPP, role limitations due to physical problems; SeF, sexual function; SF, social function.

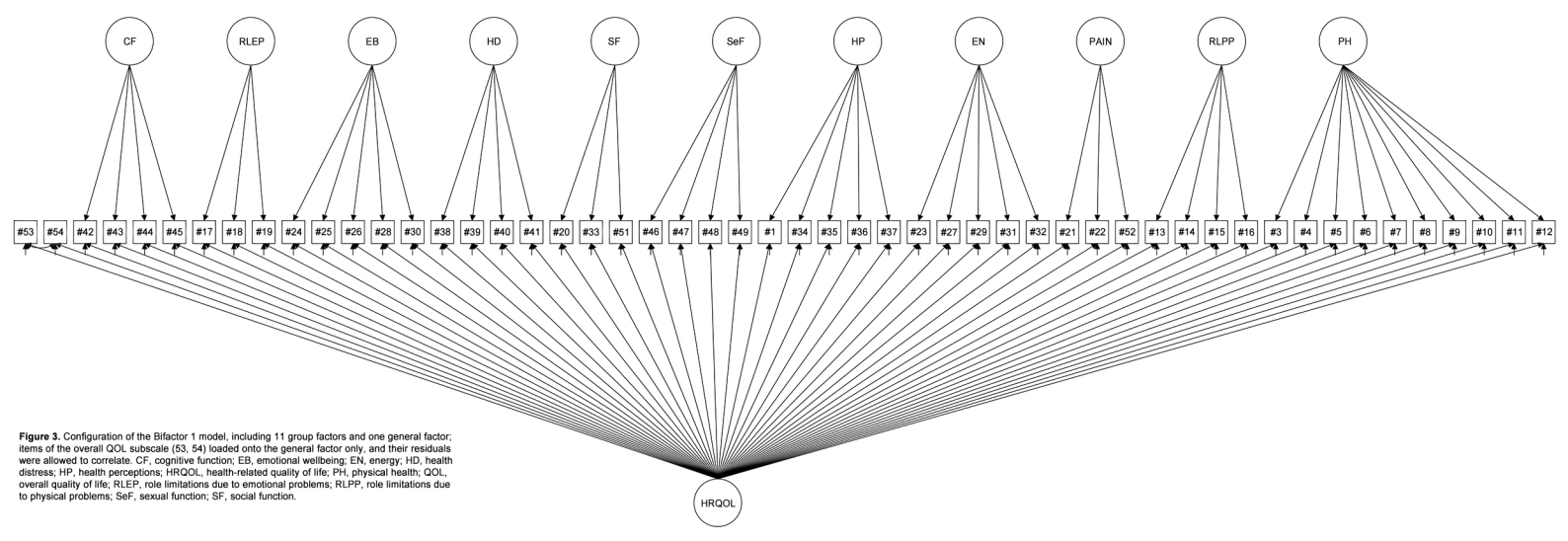


Figure 3. Configuration of the Bifactor 1 model, including 11 group factors and one general factor; items of the overall QOL subscale (53, 54) loaded onto the general factor only, and their residuals were allowed to correlate. CF, cognitive function; EB, emotional wellbeing; EN, energy; HD, health distress; HP, health perceptions; HRQOL, health-related quality of life; PH, physical health; QOL, overall quality of life; RLEP, role limitations due to emotional problems; RLPP, role limitations due to physical problems; SeF, sexual function; SF, social function.

To consider the models acceptable, we used the same criteria as in the Action 1 above: CFI >0.90; RMSEA <0.08; and SRMR <0.08 [Hu 1995, Hu 1999].

Further, Akaike Information Criterion (AIC) [Akaike 1974] and Bayesian Information Criterion (BIC) [Schwarz 1978; Stone 2009] were used for model comparisons. The AIC evaluates the goodness of fit of the model by employing the maximum value of the log-likelihood function and the total number of parameters to be estimated in the model, as follows:

$$AIC = -2L+2K$$

With the BIC, a model being more likely to generate the observed data is identified. The BIC is defined as:

$$BIC = -2L + Kln(N)$$

Here, L corresponds to the value of the log-likelihood function at its maximum value, K to the parameters in the model, and N to the sample size. The model with lower AIC and BIC values was chosen as the best model to fit the data [Akaike 1974; Schwarz 1978; Stone 2009].

To evaluate the relative strength of the general HRQOL factor to group factors, magnitude of loadings was considered (values ≥0.40 were considered satisfactory [Peipert 2019]), and explained common variance (ECV) and percentage of uncontaminated correlations (PUC) were calculated.

The ECV corresponds to the ratio of variance which is explained by general factor divided by the variance explained by the general plus group factors [Reise 2012].

The PUC corresponds to the number of unique correlations which are influenced by a single factor divided by the total number of unique correlations. A high ECV value or a moderate ECV value supplemented with a high PUC value (>0.90) indicated that data were sufficiently “unidimensional” [Reise 2012].

To judge the degree to which total raw scores reflected a common single factor, the McDonald's coefficient omega hierarchical (ω_H) was computed, as follows:

$$\omega_H = \frac{(\sum \lambda_{iGEN})^2}{(\sum \lambda_{iGEN})^2 + (\sum \lambda_{iGRP_1})^2 + (\sum \lambda_{iGRP_2})^2 \ldots + (\sum \lambda_{iGRP_p})^2 + \sum \theta_i^2}$$

Here, λ_{iGEN} is the loading of item i on the general factor; λ_{iGRP_1} is the loading of item i on the first group factor, *p* corresponds to the number of group factors, and θ_i^2 to error variance of the item i. High values indicate that the total raw score was a reliable measure of the general factor. Further, to evaluate the reliability considering all sources of common variance (general and group factors), the McDonald's coefficient omega (ω) was calculated, as follows:

$$\omega = \frac{(\sum \lambda_{iGEN})^2 + (\sum \lambda_{iGRP_1})^2 + (\sum \lambda_{iGRP_2})^2 \ldots + (\sum \lambda_{iGRP_p})^2}{(\sum \lambda_{iGEN})^2 + (\sum \lambda_{iGRP_1})^2 + (\sum \lambda_{iGRP_2})^2 \ldots + (\sum \lambda_{iGRP_p})^2 + \sum \theta_i^2}$$

Thus, both omega hierarchical and omega were also calculated for each subscale to evaluate how much subscale scores were reliable measures of the corresponding specific latent variables, once items' common variance due to the general factor was removed (ω_S), and how reliable they were considering all sources of common variance.

Finally, we used CFA to evaluate the measurement invariance of MSQOL-54 across gender (male [26%]; female [74%]), and age (using the median of 44 years old as cut-off). Recent evidence shows that average age at diagnosis across countries can vary, ranging from 20 to 50 years [Atlas of MS 2020]. We chose the median of 44 years as it is within this year range. As did in the Action 1 above and in Giordano et al. (2020), three increasingly constrained levels of measurement invariance (i.e.

configural, metric, scalar) were assessed using multi-group CFA. We used the same criteria as in the Action 1 above to assess the model fit [Giordano 2020].

All models were estimated using the software Mplus 7.0 with the maximum likelihood estimation with robust standard errors (MLR) [Muthén 2012].

RESULTS

Table 1 reports the model description and fit statistics of the CFA. The (original) two second-order factor model fitted the data quite well (RMSEA=0.055; CFI=0.888, RMRS=0.064), only the CFI index was slightly under the cut-off value. The single second-order factor model showed similar values (RMSEA=0.056; CFI=0.884, RMRS=0.068), but in terms of AIC and BIC values it was outperformed by the two second-order factor model. The bifactor model ('bifactor 1') produced good fit measures, but the solution was not robust because one item of the social function subscale (item 20 '*During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups'*) showed a negative residual variance. An inspection of the loading estimates revealed that one item (item 51 *'During the past 4 weeks, to what extent have problems with your bowel or bladder function interfered with your normal social activities with family, friends, neighbors, or groups?'*) was not a good indicator of social functioning, once parceling out the general factor. Thus, a second bifactor model ('bifactor 2') was estimated in which the three items of the social function subscale (20, 33, and 51) loaded onto the general factor only, and, to account for the group specificity of item 20 and item 33, residuals of these two items were allowed to correlate. This last model had satisfactory fit (RMSEA=0.055; CFI=0.892, RMRS=0.062), and both AIC and BIC statistic values were better (AIC=1710637; BIC=1711910) than those of the one and two second-order factor models (AIC=1711735; BIC=1712778; AIC=1711214; BIC=1712270; Table 1).

Table 1. Model description and fit statistics of confirmatory factor analysis.

Model type	AIC	BIC	RMSEA	CFI	RMRS
Two second-order factors [a]	1711214	1712270	0.055	0.888	0.064
Single second-order factor [b]	1711735	1712778	0.056	0.884	0.068
Bifactor 1[c]	1710635	1711920	0.055	0.892	0.062
Bifactor 2[d]	1710637	1711910	0.055	0.892	0.062

AIC, Akaike information criterion; BIC, Bayesian information criterion; RMSEA, root mean square error of approximation; CFI, comparative fit index; RMRS, and standardized root mean square residual (RMRS).

[a] 12 first-order factors and two second-order factors; the correlation between the two second-order factors was 0.87.
[b] 12 first-order factors and one second-order factor.
[c] 11 group factors and one general factor; residual of overall quality of life (QOL) subscale items (items 53 and 54) were allowed to correlate.
[d] 10 group factors and one general factor; items of the social function subscale (20, 32, 51) loaded onto the general factor only, and residual of the overall QOL subscale items (53, 54), and the residual of items 20 and 32 of the social function subscale were allowed to correlate.

Standardized factor loadings for the revised bifactor model are shown in Table 2. All items loaded satisfactorily on the general (HRQOL) factor (loading ≥ 0.40), the only exception being item 24 (*'Have you been a very nervous person?'*), and the four items belonging to the sexual function scale.

Loadings on the group factors were all ≥ 0.40, except for three items (item 23 *'Did you feel full of pep?'*, item 27 *'Did you have a lot of energy?'*, and item 32 *'Did you feel rested on waking in the morning?'*) of the energy subscale, two items of health perceptions (i.e. item 34 *'I seem to get sick a little easier than other people'* and item 36 *'I expect my health to get worse'*), and one of the emotional wellbeing subscale (item 26 *'Have you felt calm and peaceful'*).

ECV value was 0.51 (indicating that 51% of the common variance was due to the general HRQOL factor) and PUC was 0.92, denoting that the data were sufficiently 'unidimensional'.

Omega value for the total raw score was 0.98, suggesting that the reliability considering all sources of common variance (general factor and group factors) was very high. Moreover, omega hierarchical value of the general factor was 0.87, indicating that the total raw score was a reliability measure of

the general HRQOL factor, and that a high proportion of the reliable variance (0.87/0.98 = 89%) in the total raw score could be accounted for by the general factor.

Table 2. Standardized factor loadings in the bifactor model.

Scales	Items	Factor loading	
		General HRQOL factor	Group factor
Physical function	3. Vigorous activities	0.553	0.445
	4. Moderate activities	0.594	0.622
	5. Lift, carry groceries	0.554	0.620
	6. Climb several flights	0.569	0.665
	7. Climb one flight	0.533	0.695
	8. Bend, kneel	0.551	0.594
	9. Walk mile	0.555	0.669
	10. Walk several blocks	0.526	0.734
	11. Walk one block	0.488	0.726
	12. Bath, Dress	0.461	0.523
Role limitations due to physical problems	13. Cut down time	0.542	0.537
	14. Accomplished less	0.571	0.571
	15. Limited in kind	0.581	0.645
	16. Had difficulty	0.594	0.586
Role limitations due to emotional problems	17. Cut down time	0.504	0.646
	18. Accomplished less	0.515	0.699
	19. Not careful	0.509	0.592
Bodily pain	21. Pain magnitude	0.575	0.702
	22. Pain interfere with work	0.611	0.653
	52. Pain interfere with enjoyment	0.601	0.652
Emotional wellbeing	24. Nervous person	**0.371**	0.531
	25. Down in dumps	0.561	0.585
	26. Peaceful	0.562	**0.369**
	28. Blue/Sad	0.594	0.592
	30. Happy	0.535	0.432
Energy	23. Pep/life	0.713	**0.206**
	27. Energy	0.717	**0.245**
	29. Worn out	0.624	0.546
	31. Tired	0.620	0.602
	32. Rested on walking in the morning	0.519	**0.281**
Health perceptions	1. EVGFP rating	0.638	0.452
	34. Sick easier	0.417	**0.269**
	35. As healthy	0.463	0.575
	36. Health to get worse	0.450	**0.233**
	37. Health excellent	0.590	0.659
Cognitive function	42. Concentration and thinking	0.591	0.710
	43. Sustained attention	0.576	0.700
	44. Memory	0.467	0.708
	45. Others note troubles with memory/concentration	0.436	0.564
Health distress	38. Discouraged	0.729	0.508
	39. Frustrated	0.712	0.544
	40. Worried for life	0.624	0.543
	41. Weighed down	0.694	0.563
Sexual function	46. Lack if sexual interest	**0.346**	0.684
	47. Erection/Lubrication	**0.299**	0.758
	48. Orgasm	**0.348**	0.724
	49. Satisfy sexual partner	**0.378**	0.656
Social function	20. Social extent, physical health	0.737	-
	33. Social time	0.758	-
	51. Social extent, bowel or bladder	0.505	-
Overall quality of life	53. 0-10 NRS rating	0.735	-
	54. TUMMMPO rating	0.685	-

EVGFP, Excellent, Very good, Good, Fair, Poor. HRQOL, health-related quality of life. NRS, Numeric Rating Scale. TUMMMPO, Terrible, Unhappy, Mostly dissatisfied, Mixed – about equally satisfied and dissatisfied, Pleased, Delighted.
Correlations between residuals: 0.524 (items 53 and 54); 0.411 (items 20 and 33).
Coefficients <0.40 are reported in bold; all the loadings are statistically significant at $p < 0.001$.

As shown in Table 3, for the majority of the subscales, omega hierarchical value (ω_S) was around 0.50, whereas it was very low (≤ 0.35) for three subscales (i.e. energy, health perceptions, and health distress) – meaning that summed scores of items belonging to these subscales were not a reliable measure of their respective domain latent variable once the general HRQOL was taken into account – and it was high (0.70) for sexual function subscale. For the latter subscale, it seems that the specific group factor accounted for more variance than the general factor, indicating that items belonging to this subscale were more likely to reflect a specific domain of HRQOL (related to sexual function) than a common general construct of HRQOL.

Table 3. Omega statistics for the MSQOL-54 total and subscales scores.

Subscale	**No. of items**	**ω**	**ω_S**
Physical function	10	0.96	0.55
Role limitations due to physical problems	4	0.89	0.46
Role limitations due to emotional problems	3	0.86	0.53
Bodily pain	3	0.92	0.52
Emotional wellbeing	5	0.85	0.41
Energy	5	0.86	0.22
Health perceptions	5	0.82	0.35
Cognitive function	4	0.91	0.57
Health distress	4	0.93	0.35
Sexual function	4	0.87	0.70

Note. ω = scores reliability considering all sources of common variance (the general and the group factor); ω_S (omega hierarchical subscale) = scores reliability considering only the common variance due to the group factor, that is the reliability of subscales scores, controlling for the effects of the general factor.

Measurement invariance

Based on the 'bifactor 2' model above, we carried out further analyses to check whether the bifactor solution was invariant across gender and age.

First, the model was estimated to evaluate the measurement invariance of MSQOL-54 across gender (Table 4, upper part). Results showed that the model produced an acceptable fit for configural invariance (RMSEA=0.055; CFI=0.892; SRMR=0.063). Considering the model where loadings were imposed to be identical across gender, indices of fit were satisfactory, and worsening of the unrestrained model was insignificant (ΔRMSEA < 0.001; ΔCFI= − 0.006; ΔSRMR = 0.008), hence providing evidence of metric invariance. With regard to the scalar invariance (i.e. intercepts and loadings imposed to be invariant across groups), the model fitted the data well (RMSEA=0.054; CFI=0.885; SRMR=0.063). Finally, examining the variations in fit indices when compared with the metric invariance model, cut-off values were met, supporting the scalar invariance.

Second, the model was estimated to evaluate the measurement invariance of MSQOL-54 across age (using the median of 44 years as cut-off) (Table 4, bottom part). Here, the results showed that the model produced acceptable fit for configural invariance (RMSEA=0.054; CFI=0.893; SRMR=0.059), metric invariance (RMSEA=0.054; CFI=0.889; SRMR=0.063), and scalar invariance (RMSEA=0.054; CFI=0.885; SRMR=0.063). All the changes in fit indices across the models were satisfactory.

Table 4. Measurement invariance of MSQOL-54 across gender and age.

	χ^2(df)[a]	RMSEA	CFI	SRMR	Δ RMSEA	Δ CFI	Δ SRMR
Male	4658.7 (1225)	0.054	0.891	0.063			
Female	11160.1 (1225)	0.055	0.893	0.062			
Configural invariance	15829.7 (2450)	0.055	0.892	0.063			
Metric invariance	16126.2 (2538)	0.054	0.891	0.065	-0.001	-0.001	0.002
Scalar invariance	16598.8 (2579)	0.055	0.887	0.065	0.001	-0.004	0.000
Adults <44 years old	7253.1 (1225)	0.053	0.890	0.056			
Adults ≥ 44 years old	7811.9 (1225)	0.054	0.895	0.061			
Configural invariance	15047.9 (2450)	0.054	0.893	0.059			
Metric invariance	15588.1 (2538)	0.054	0.889	0.063	0.000	-0.004	0.004
Scalar invariance	16084.7 (2579)	0.054	0.885	0.063	0.000	-0.002	-0.004

CFI, comparative fit index; df, degrees of freedom; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

[a] χ^2 p-values are all <0.001.

DISCUSSION

The bifactor model is particularly useful when it is intended to acknowledge multidimensionality and, at the same time, take account of a single general construct [Reise 2012], as the HRQOL related to MS is. As far as we know, this was the first study applying the bifactor model to the MSQOL-54 in a large international database of MS patients.

The bifactor model with one general HRQOL factor and 10 specific group factors achieved acceptable fit and outperformed both the original two second-order factor model and the single second-order factor model. Also, our findings supported measurement invariance of the questionnaire across age and gender, suggesting that it has the same meaning across these socio-demographic variables, and that patients having the same ratings on MSQOL-54 general or domain factors would attain the identical value on the observed variable, regardless of sub-group membership.

Generally, the factor loadings were substantially high both on the general and the group factors, and the ECV was about 50%, indicating that MSQOL-54 items contribute to essentially the same

extent to both the general HRQOL factor and to the group factors. Despite this, the data can be deemed sufficiently 'unidimensional', because the MSQOL-54 consists of several subscales composed of few items each, and this implies that the vast majority of correlations between items (PUC=92%) reflect general factor variance only. Furthermore, the satisfactory value of the coefficient omega hierarchical indicated that the total raw score is a reliable measure of the general HRQOL latent variable. Taken together, all these results support the hypothesis that the MSQOL-54 has a sufficient 'unidimensional' structure, and thus it is appropriate to calculate a total HRQOL score.

Among the 52 items analyzed in the study – it is noteworthy to remember that items 2 and 50 were excluded from the analysis as they are single items – the weaker indicators of the general HRQOL dimension were the four items of the sexual function subscale. Considering the omega hierarchical value, the sexual function subscale is more likely to reflect a specific domain of HRQOL (namely related to sexual function) than a common general construct of HRQOL. In fact, this is the only subscale that showed an omega hierarchical value ≥ 0.70.

Another issue derives from the social function subscale. The three items of this subscale loaded onto the general factor only because one of them (item 51, dealing with bowel or bladder) was not a good indicator of social functioning, and a group factor needs at least three items to be identified. Thus, it was not possible to evaluate the contribution of the relative group factor.

This study has important implications for clinical practice and research. For clinical practice, it could be crucial to provide health professionals, MS patients with feedback using a single HRQOL total score, as well as with subscale scores, to add granularity. The total HRQOL score could be useful also to identify patient subgroups in order to deliver personalized interventions addressing, for example, self-efficacy or resilience. On the other hand, for researchers, it could be easier to calculate and interpret a unique total HRQOL score, when using such measure in clinical trials or other research

studies. Moreover, the present results can be a stimulus for future research aimed at revising the MSQOL-54 questionnaire. Specifically, our findings highlight the need to enlarge the number of items measuring the social function subscale, because one of the three items of this subscale was not a good indicator. Furthermore, we suggest revising the sexual function subscale items by broadening the content domain so as to include also intimacy and sexual pleasure, as three of the four items from this subscale originated from Medical Outcomes Study sexuality functioning scale which focus on performance indicators [Sherbourne 1992].

In the present study there were a number of limitations, some of which are reported elsewhere [Giordano 2020]. This secondary analysis was carried out in a large cross-sectional international MS database and should be confirmed in an independent sample, using a prospective longitudinal design.

CONCLUSION

To conclude, this study adds new knowledge to the factorial structure of the MSQOL-54, in that a bifactor model fits the data well, outperforming the two second-order models. Therefore, it is appropriate to calculate a total HRQOL score, including all the original subscales/domains. Based on these results, in future research, items should be calibrated using IRT in order to assess whether a multidimensional CAT version of the MSQOL-54 is feasible. Further work to integrate/revise selected items is needed.

ACTION 3 - Development of the multidimensional CAT version of the MSQOL-54

Preface

The Action 2 presented above provided evidence that a bifactor model fits the data well suggesting that an overall HRQOL score can be computed using the MSQOL-54. By incorporating the multidimensionality produced by sampling of items from pre-specified group factors, the bifactor model is particularly suited to the measurement of multidimensional HRQOL and their sub-components, at the same time providing a single overall HRQOL score. This eases the subsequent CAT and reduces the number of items required for adaptive test.

The present Action 3 describes the procedures followed to calibrate items according to the multidimensional IRT analysis using a bifactor model, as a key premise to apply multidimensional CAT to the MSQOL-54, and assess its performance, in comparison to the fixed-length questionnaire. A paper including the results of this Action is being prepared.

Rationale

As HRQOL assessment includes assessment of patients' physical, psychological, and social functioning, together with the impact of disease and treatment on their abilities and daily functioning, usually HRQOL questionnaires are rather long. As described in the chapter 1 above, CAT can provide patients with individualized items, maximizing the information obtained, shortening the questionnaire length, and thus reducing patient and clinician burden [Wainer 2000]. By taking into account correlations between domains, MCAT may be a more efficient approach to assess HRQOL [van der Linden 2010].

The MSQOL-54 is the most used MS-specific HRQOL inventory [Solari 2005], but its application in clinical setting is restricted by its length, patients’ and clinicians' burden. Moreover, the

questionnaire has limitations such as its length, a possible floor effect for physical function scale, and a high number of missing answers for 'sexual function' and 'satisfaction with sexual function scales'.

Recent evidence shows that MCATs applied to shorten fixed-length available HRQOL questionnaires are scarce [Michel 2016; Rebollo 2010].

The aim of the present Action was to develop MCAT for MSQOL-54 and to investigate its performance in terms of item (i.e. question) reduction and preservation of a reliable score estimate.

METHODS

Patient sample

To perform the present secondary analysis, we used the same data described in the Actions 1 and 2 above, and in a recent publication [Giordano 2020].

MIRT modelling with bifactor model

In the present work, before implementing CAT, we calibrated an item bank using a bifactor IRT model for graded response data [Gibbons 2007], which relates properties of the test items (e.g., their difficulty and discrimination) to the "ability" (or other trait) of the examinee, following the protocol reported by the PROMIS investigators [Reeve 2007]. As described in the Action 2 above, in the bifactor model, items load on a general factor and only on one group factor. Moreover, the general and group factors are all uncorrelated to each other [Reise 2010].

According to the bifactor factor structure of the MSQOL-54 resulted in the Action 2 above (i.e. 'Bifactor 2', Table 1, Action 2), items 2 and 50 were not included in this model, because they are single items. Further, five items loaded onto general factor only: items 53 and 54 forming the overall

QOL subscale, because the bifactor model needs each group factor to be composed of at least three items to be identified; and the three items of the social function subscale (20, 33, and 51).

In order to assess the local independence assumption of the items, the Yen's Q3 index has been calculated [Yen 1984]. The Q3 index is calculated for every item pair (i,j) and corresponds to the correlation between item residuals after fitting the model. These item residuals are differences between the observed responses of the individual item and the response reproduced by the model. We considered item residual correlations above 0.20 to be indicative of local dependence between items [Yen 1984]. Items with local dependence were removed from the subsequent analyses [Reeve 2007]. Missing data were handled by using a full information maximum likelihood method of estimation.

True θs and their MCAT estimation

Each examinee latent score (θ) was randomly drawn from a multivariate normal (MVN) distribution with MVN (0, I), without any correlations among θ_S, as the latent traits were assumed to be orthogonal. These true θ_s were generated with the ''mvtnorm'' package [Genz & Bretz 2009] in R for the bifactor model, with one general factor and ten group factors.

Given the observed response vector for each examinee, individual latent trait scores for the general and group factors are estimated via the multidimensional maximum a posteriori (MAP) estimator [Gibbons 2007].

Item selection for MCAT

The bifactor IRT items parameters and a matrix of simulated item responses derived from a multivariate normal distribution were simultaneously processed. Each CAT begins with an initial θ estimate of 0.0, and the Fisher information item selection method was used. The item with the

greatest item information with the initial latent value was chosen as the first item; the items which maximized the determinant of the information matrix at the updated theta estimates produced from responses to the previous items were then selected after (D-Optimality criteria; Seo and Weiss 2015).

Stopping rule

Finally, the CAT terminated using a fixed standard error of the θ estimate (SEM), allowing the number of items to vary across examinees.

CAT simulations

We run a simulation study with the two standard errors set up at 0.40 and 0.32 on general HRQOL factor, as the stopping rule. We chose these values as they correspond to Cronbach alpha thresholds of 0.85, and 0.90, respectively. In addition, these thresholds were employed in other studies in the HRQOL field [Loe 2017; Geerards 2019].

Performance of the CAT was assessed by calculating the root mean square error (RMSE), bias, and the mean number (minimum-maximum) of items administered. RMSE and bias were determined by comparing CAT estimated scores with simulated true scores.

RMSE, and bias were calculated as follows:

$$RMSE\ (\hat{\theta}) = \sqrt{\frac{\sum_{J=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}}$$

$$Bias\ \left(\hat{\theta}\right) = \frac{\sum_{J=1}^{N}(\hat{\theta}_j - \theta_j)}{N}$$

Here, $\hat{\theta}_j$represents estimated θ level for the j^{th} examinee for each research condition tested, θ_i indicates the true θ level for each examinee, as defined above, and *N* is the number of examinees

[Yen 1987; Harwell 1991; Gao 2005]. A low RMSE value indicates a more accurate measurement [Sunderland 2020].

Software

We performed the analysis using R (version 3.4.3). We modeled the responses to the MSQOL-54 items using the bifactor IRT model with *mirt* package [Chalmers 2012].

We developed the program with the CAT algorithms with the package *mirtCAT* in R [Chalmers, 2016].

RESULTS

Bifactor IRT modeling

We started the analysis by assessing whether items also met the assumption of local independence. Such local dependency was apparently (i.e. residual correlations >0.2) between items 5 and 10, 30 and 54, 9 and 10, 6 and 7, 4 and 5, 10 and 11, 44 and 45, 20 and 33, 29 and 31 (Table 1).

In particular, items 30 (*Have you been a happy person?*) and 54 (*Which best describes how you feel about your life as a whole? Terrible, Unhappy, Mostly dissatisfied, Mixed – about equally satisfied and dissatisfied, Pleased, Delighted*) have similar content, as for items 20 (*During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups'*) and 33 *(How much of the time as your physical health or emotional problems interfered with your social activities, like visiting with friends, relatives, etc.?)*.

Further, items 29 (*Did you feel worn out?*) and 31 (*Did you feel tired?)* have similar stem.

Items 4 *(moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, etc.)* and 5 (*lifting or carrying groceries*) have similar content and are presented sequentially, as for items 44

(Have you have trouble with your memory) and 45 *(Have others, such as family members or friends, noticed that you have trouble with your memory or problems with your concentration?*); and for items 53 *(Overall, how would you rate your own quality of life?)* and 54 ((*Which best describes how you feel about your life as a whole? Terrible, Unhappy, Mostly dissatisfied, Mixed – about equally satisfied and dissatisfied, Pleased, Delighted*).

Finally, items 9 (*walking more than a mile*) and 10 (*walking several blocks*) have similar stem, content, and are presented sequentially, as for items 6 *(climbing several flights of stairs)* and 7 *(climbing one flight of stairs)*; and for items 10 (*walking several blocks*) and 11 (*walking one block*). Thus, we decided to remove 7 items from the subsequent analysis (i.e. 10, 54, 6, 5, 45, 20, 31).

Table 1. Item residual correlations and items which were removed.

Item no.	Item no.	Yen's Q3 value	Item removed
5	10	0.22	10
30	54	0.25	54
9	10	0.26	-
6	7	0.27	6
4	5	0.30	5
10	11	0.31	-
44	45	0.37	45
20	32	0.39	20
29	31	0.47	31
53	54	0.51	-

CAT simulations

The matrix of item parameter estimates from the bifactor IRT calibration including 45 items, and the matrix of simulated item responses derived from a multivariate normal distribution were processed. When the stopping rule was set at SEM=0.40, the mean number of items administered was 10 (7-45) (a reduction of 78%), resulting in bias= 0.10, and RMSE=0.41 compared with scores estimated using all 45 items.

When the stopping rule was set at SEM=0.32, the mean number of items administered was 26 (18-45) (a reduction of 58%) resulting in bias= 0.0008, and RMSE=0.33 compared with scores estimated using all 45 items (Table 2).

Table 2. MSQOL-54 MCAT simulation of CAT algorithm.

Measure	SEM=0.40	SEM=0.32
Mean item administered (range)	10 (7-45)	26 (18-45)
Bias	0.10	0.0008
RMSE	0.41	0.33

RMSE, root mean square error. SEM, standard error of measurement.

DISCUSSION

In the present Action, we firstly calibrated items by applying MIRT using bifactor model. Then, we performed CAT simulations using two different values for the standard error of estimates as the stopping rule (i.e. 0.40, and 0.32, respectively), and assessed their performance. Importantly, findings from the CAT simulations showed that the CAT administration proved to be parsimonious. In particular, the mean number of administered items was 10 (7-45), and 26 (18-45), when the SEM was fixed at 0.40 and 0.32, respectively. Error indices (i.e. bias and RMSE) showed lower values for the second simulation in comparison to the first one, suggesting that the second was more accurate than the first. In fact, questionnaires with many items may yield high levels of scale reliability.

To the best of our knowledge, this is the first application of MCAT to the MSQOL-54. Research in this field is sparse and there are a few examples in MS and neurology fields reporting interesting results using other instruments, such as the MusiQOL [Michel 2016], and the Neuro-QOL [Healy 2019].

As far as the methods of the MCAT are concerned, a few issues should be discussed. First, to estimate latent traits, we chose MAP algorithm rather than the EAP, as it is more appropriate. In fact, Yao (2014) found that MAP provides better precision than maximum likelihood, and performs as well as the EAP. In addition, Chalmers (2012) suggested that MAP should be used instead of EAP when higher dimensional models are considered, like the bifactor model is in our study.

Second, to select subsequent items, we chose D-rule instead of other rules reported in the literature [Seo & Weiss 2015], as D-rule improves theta estimates with regard to a general factor [Seo & Weiss 2015], as the HRQOL general factor is.

Third, there are several potential stopping rules to apply when developing MCAT simulations, as for example the amount of time, or the fixed-length rule, that could be more compatible with clinical practice. We chose the two different standard errors of 0.32 and 0.40 as stopping rule, as it is well-recognized in the literature that are equivalent to levels of reliability higher than 0.80 [Loe 2017].

Our study has however some limitations. First, we did use a fixed-length questionnaire (i.e. the MSQOL-54) and perform CAT simulations. This issue is well-recognized in the HRQOL literature [Michel 2016], and, as discussed also in the Action 2 above, further work should be conducted to add/integrate/revise items of the MSQOL-54, in order to make the calibration and MCAT performance even more efficient. In fact, although the MSQOL-54 was constructed to be a comprehensive measure of HRQOL in MS, it was developed in 1995, and it was suggested that investigators should perform such item 'seeding' at a certain time to maintain and renew item banks [Wainer 2000]. Notably, this is even more important and challenging in the HRQOL field where questionnaires are usually multidimensional.

Another limitation is that, in the MCAT simulations, we preferred to use a matrix of simulated item responses. A few disadvantages of these simulations should be acknowledged, such as that they take time and effort to produce and that results may seem less relevant as they are obtained in the

improbable situation of the data perfectly fitting the calibrated model. On the other hand, results from simulations using 'real-data' are readily available and are more convincing to the audience. However, when using 'real data' the true latent trait levels are not available and results cannot be generalized to other patient groups [Smits 2018].

Finally, although other simulations have been planned in advance with the objective to work on different estimation models and item selection criteria, due to the current COVID-19 pandemic it was not possible to use the dedicated PC available in the University Lab and perform such analyses. As they are preliminary, these results should be considered with caution.

GENERAL DISCUSSION

The present thesis, which is part of an international collaboration between Italian and Australian investigators, aimed to develop a MCAT version of the MSQOL-54. Results show that the computerized adaptive version of the MSQOL-54 is feasible, thus reducing item number and patient, clinician, and researcher burden.

To the best of our knowledge, this the first example of a newly-developed CAT version of the MSQOL-54. As such, this research has many strengths which should be mentioned. First, we used responses from a large international sample of 3669 MS patients which consists of Italian and English language versions of the MSQOL-54. In the Action 1 above, we assessed the measurement invariance of the two language versions and found that the questionnaire has the same meaning across languages [Giordano 2020]. However, we are aware that the recruitment strategies adopted to gather Italian and English data were different, as reported also elsewhere [Giordano 20020], and our data could not be generalizable to other MS patient samples. Based on the Action 1 results, we merged the responses and obtained a unique dataset. Then in the Action 2, we applied bifactor modeling to the MSQOL-54 and provided new evidence on the dimensionality of the questionnaire, in that a general HRQOL score was viable. This could have a potential impact on clinical practice in that this HRQOL general score can be used for measurement, for example to assess treatment response / modify treatment plan on an underlying continuous scale of measurement. Based on these findings, in the Action 3, we calibrated items using MIRT with bifactor model, and demonstrated that the MCAT version of the MSQOL-54 is feasible, with a reduction of about 70% of the items.

Based on these results, a number of next steps could be performed in the future. As reported above, after working on adding/integrating/revising items of the MSQOL-54, validation studies using an

independent MS sample - rather than simulation of existing data, should be prospectively conducted, by including other clinical and socio-demographic variables (e.g. levels of education, employment, and disease forms) as well as other relevant PROs. This could be done in order to explore better the CAT performance and the external validity of the computerized adaptive version of the MSQOL-54. The same validation studies could be performed using a longitudinal design, so as to assess over time other important psychometric properties, such as for example the test-retest reliability or the sensitivity to change. In such studies, among the others, the Concerto testing platform (University of Cambridge, The Psychometrics Center, Cambridge, United Kingdom), a flexible open-source tool could be used to deploy CAT to the patients, using also mobile devices [Harrison 2020].

Overall, this study has also important implications for clinical practice and research. By reducing item numbers and tailoring items to the individual patient, thus improving the efficiency and precision of the instrument, the MCAT-MSQOL-54 would provide patients, clinicians, and researchers with immediate feedback. This will increase accuracy, make the test interpretable, and shorten the time spent for questionnaire administration, reducing patient burden. In our two simulations a reduction of 78%-58% of administered items was reported. This could have consequently much impact on clinical practice, where time is at premium. Although preliminary, these results could have also an impact of patient-physician relationship as well as shared decision making. In fact, incorporating patient's perspective is crucial to improve outcomes of care and it is, at the same time, a key component of patient-centered care [Heesen 2011].

Besides enriching the patient-physician relationship, the MCAT-MSQOL-54 could be employed also at the group level data. Indeed, it could be also integrated in the electronic health records, as well as in the MS registries, both at the national [Trojano 2019] and international levels. A few examples

of clinician-based registries are the EDMUS (European Database for Multiple Sclerosis), the MS-COSTAR (Multiple Sclerosis–Computer Storage Ambulatory Records) [Confavreaux 1995], and the Danish Registry [Koch-Henriksen 2001]. Further, examples of patient-reported registries collecting QOL data are the Sonya Slifka Longitudinal MS Study [Minden 2006]; NARCOMS (North American Research Committee on Multiple Sclerosis) database [Consortium of MS Centers].

Further, another novel method to incorporate such MCAT-MSQOL-54 into practice could be the patient portals websites. These portals are generally linked also to electronic health records, allowing the patients to monitor their health. A few interesting examples of this potential integration between the two systems could be found in mental health setting in US [Gibbons 2012; Gibbons 2013].

With the objective to make the information immediately available to patients, such portals may represent the next step further to integrate PROs into clinical practice, thus improving quality of care.

CONCLUSION

To conclude, the results of the present thesis, included in the ongoing international collaborative project between Italian and Australian investigators, indicated that the MCAT version of the MSQOL-54 is feasible, providing notable item reduction and reducing patient burden, while preserving high accuracy levels. Further work to revise the original MSQOL-54 item bank, and improve CAT deployment should be conducted. The MCAT-MSQOL-54 version could be used in clinical practice and research.

References

Aaronson NK, Acquadro C, Alonso J, et al. International quality of life assessment (IQOLA) project. *Qual Life Res* 1992; 1: 349–51.

Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993; 85(5): 365-76.

Acquadro C, Lafortune L, Mear I. Quality of life in multiple sclerosis: translation in French Canadian of the MSQoL-54. *Health Qual Life Outcomes* 2003; 1: 70.

Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974; 19: 716–23.

Allen DD, Ni P, Haley SM. Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. *Disabil Rehabil* 2008; 30: 479-84.

Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; 43: 561–73. [a]

Andrich D. Application of a psychometric rating model to ordered categories, which are scored with successive integers. *Applied Psychological Measurement* 1978; 2: 581–594. [b]

Aronson KJ. Quality of life among persons with multiple sclerosis and their caregivers. *Neurol* 1997; 48: 74-80.

Atlas of MS 2020. https://www.msif.org/wp-content/uploads/2020/10/Atlas-3rd-Edition-Epidemiology-report-EN-updated-30-9-20.pdf.

Bassi M, Falautano M, Cilia S, et al. (2016). Illness Perception and Well-Being Among Persons with Multiple Sclerosis and their caregivers. *J Clinical Psychology Medical Settings* 2016; 23: 33-52.

Bassi M, Falautano M, Cilia S, et al. The coexistence of well- and ill-being in persons with multiple sclerosis, their caregivers and health professionals. *J Neurol Sci* 2014; 337: 67-73.

Benito-León J, Morales JM, Rivera-Navarro A, Mitchell AJ. A review about the impact of multiple sclerosis on health-related quality of life. *Disabil Rehabil* 2003; 25: 1291–303.

Bentler PM (1995). EQS structural equations program manual. Encino, CA: Multivariate Software, Inc.

Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981; 19(8): 787-805.

Birnbaum A. (1968) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R., Eds., Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, pp. 397-479.

Bjorner JB, Kosinski M, Ware JE Jr. Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Qual Life Res* 2003; 12: 981-1002.

Bludworth JL, Tracey TJG, Glidden-Tracey C. The bi-level structure of the Outcome Questionnaire-45. *Psychological Assessment* 2010; 22: 350–5.

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 1981; 46: 443-59.

Bock RD, Gibbons R, Muraki E. Full-information item factor analysis. *Applied Psychological Measurement* 1988; 12: 261-80.

Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement* 1982: 6: 431-44.

Boer D, Hanke K, He J. On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests. *Journal Cross-Cultural Psychology 2018;* 49(5): 713–34.

Brichetto G, Zaratin P. Measuring outcomes that matter most to people with multiple sclerosis: the role of patient-reported outcomes. *Curr Opin Neurol* 2020; 33(3): 295-9.

Brook R, Ware J, Davies-Avery A, et al. *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume VIII, Overview*. Santa Monica, CA: The Rand Corporation; 1979.

Brooks R, EuroQol group. EuroQol: the current status of play. *Health Policy* 1996; 37: 53-72.

Brouwer D, Meijer RR, Weekers AM, Baneke JJ. On the dimensionality of the Dispositional Hope Scale. *Psychological Assessment* 2008; 20: 310–5.

Browne MW, Cudeck R (1993). Alternative ways of assessing model fit. In Bollen KA & Long JS (Eds.). Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.

Bruce JM, Hancock LM, Arnett P, Lynch S. Treatment adherence in multiple sclerosis: association with emotional status, personality, and cognition. *J Behav Med* 2010; 33: 219–27.

Brunet DG, Hopman WM, Singer MA, Edgar CM, MacKenzie TA. Measurement of health-related quality of life in multiple sclerosis patients. *Can J Neurol Sci* 1996; 23: 99-103.

Burden of illness of multiple sclerosis: Part II: Quality of life. The Canadian Burden of Illness Study Group. *Can J Neurol Sci* 1998; 25: 31-38.

Byrne BM, Shavelson R J, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 1989; 105(3): 456.

Calman KC. Quality of life in cancer patients: an hypothesis. *J Med Ethics* 1984; 10: 124-7.

Cano SJ, Hobart JC. The problem with health measurement. *Patient Prefer Adherence* 2011; 5: 279-90.

Cella DF, Dineen MA, Arnason B, et al. Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. *Neurol* 1996; 47: 129-39.

Cella DF, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *J Clin Oncol* 1993; 11(3): 570-9.

Chalmers RP. Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software* 2016; 71. DOI: 10.18637/jss.v071.i05.

Chalmers RP. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* 2012; 48(6): 1–29.

Chamot E, Kister I, Cutter GR. Item response theory-based measure of global disability in multiple sclerosis derived from the Performance Scales and related items. *BMC Neurol* 2014; 3: 192. doi: 10.1186/s12883-014-0192-1.

Chen FF, West SG, Sousa KH. A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research 2006;* 41(2): 189–225.

Chen FF. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal* 2007; 14(3): 464–504.

Chilcot J, Norton S, Kelly ME, Moss-Morris R. The Chalder Fatigue Questionnaire is a valid and reliable measure of perceived fatigue severity in multiple sclerosis. *Mult Scler 2*016; 22: 677-84.

Christensen KB, Makransky G, Horton MC. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch model using Residual Correlations. *Applied Psychological Measurement* 2017; 41: 178-94. https://doi.org/10.1177/0146621616677520.

Chung H, Kim J, Askew RL, Jones SM, Cook KF, Amtmann D. Assessing measurement invariance of three depression scales between neurologic samples and community samples. *Qual Life Res* 2015; 24: 1829-34.

Chung H, Kim J, Park R, Bamer AM, Bocell FD, Amtmann D. Testing the measurement invariance of the University of Washington Self-Efficacy Scale short form across four diagnostic subgroups. *Qual Life Res* 2016; 25: 2559-64.

Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. *Ann Acad Med Singap* 1994; 23(2): 129-38.

Committee for Medicinal Products for Human Use (CHMP). EMA/CHMP/771815/2011, Rev. 2. Guideline on clinical investigation of medicinal products for the treatment of Multiple Sclerosis. Available at https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-investigation-medicinal-products-treatment-multiple-sclerosis_en-0.pdf.

Compston A, Coles A. Multiple sclerosis. *Lancet* 2008; 372: 1502–17.

Compston A, McDonald I, Noseworthy J, et al. *McAlpine's Multiple Sclerosis*, 4th edn. Edinburgh, UK: Churchill Livingstone Elsevier, 2006.

Confavreux C, Paty DW: Current status of computerization of multiple sclerosis clinical data for research in Europe and North America: the EDMUS/MS-COSTAR connection. European Database for Multiple Sclerosis. Multiple Sclerosis-Computed Stored Ambulatory Record. *Neurology* 1995; 45: 573–6.

Consortium of Multiple Sclerosis Centers: NARCOMS Multiple Sclerosis Registry. Available at https://www.narcoms.org.

Cox SD, Pakenham KI. Confirmatory factor analysis and invariance testing of the Young Carer of Parents Inventory (YCOPI). *Rehabilitation Psychology* 2014; 59: 439-52.

de Haes JC, van Knippenberg FC, Neijt JP. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. *Br J Cancer* 1990; 62(6): 1034-8.

Degenhardt A, Ramagopalan SV, Scalfari A, Ebers GC. Clinical prognostic factors in multiple sclerosis: a natural history review. *Nat Rev Neurol* 2009; 5: 672–82.

Dennison L, Brown M, Kirby S, Galea I. Do people with multiple sclerosis want to know their prognosis? A UK nationwide study. *PLoS ONE* 2018; 13: e0193407.

Dennison L, McCloy Smith E, Bradbury K, Galea I. How do people with multiple sclerosis experience prognostic uncertainty and prognosis communication? A qualitative study. *PLoS ONE* 2016; 11: e0158982.

Devinsky O, Vickrey BG, Cramer J, et al. Development of the quality of life in epilepsy inventory. *Epilepsia* 1995; 36(11): 1089-104.

Ebers GC, Traboulsee A, Li D, et al. Analysis of clinical outcomes according to original treatment groups 16 years after the pivotal IFNB- 1b trial. *J Neurol Neurosurg Psychiatry* 2010; 81(8): 907–12.

El Alaoui Taoussi K, Ait Ben Haddou E, Benomar A, Abouqal R, Yahyaoui M. Quality of life and multiple sclerosis: Arabic language translation and transcultural. adaptation of MSQOL-54. *Revue Neurologique* 2012; 168: 444-9.

Embretson SE, Reise SP (eds). *Item response theory for psychologists*. Lorence Erlbaum Associates, 2000.

EuroQol. https://euroqol.org.

Fayers MF, Machin D (eds). *Quality of life: the assessment, analysis and reporting of patient-reported outcomes*. Hoboken, New Jersey, US: Wiley Blackwell; 2016.

Fischer JS, LaRocca NG, Miller DM, Ritvo PG, Andrews H, Paty D. Recent developments in the assessment of quality of life in multiple sclerosis (MS). *Mult Scler* 1999; 5: 251-9.

Ford HL, Gerry E, Tennant A, Whalley D, Haigh R, Johnson MH. Developing a disease-specific quality of life measure for people with multiple sclerosis. *Clin Rehabil* 2001; 15: 247-58.

Freedman M, Hughes B, Mikol D, et al. Efficacy of disease-modifying therapies in relapsing remitting multiple sclerosis: a systematic comparison. *Eur J Neurol* 2008; 60: 1–11.

Füvesi J, Bencsik K, Benedek K, et al. Cross-cultural adaptation and validation of the 'Multiple Sclerosis Quality of Life Instrument' in Hungarian. *Mult Scler* 2008; 14: 391-8.

Gao F, Chen L. Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three- parameter logistic model. *Applied Measurement in Education* 2005; 18(4): 351–80.

GBD 2016 Multiple Sclerosis Collaborators. Global, regional, and national burden of multiple sclerosis 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurology* 2019; 18(3): 269-85.

Geerards D, Klassen AF, Hoogbergen MM, et al. Streamlining the Assessment of Patient-Reported Outcomes in Weight Loss and Body Contouring Patients: Applying Computerized Adaptive Testing to the BODY-Q. *Plast Reconstr Surg* 2019; 143(5): 946e-955e. doi: 10.1097/PRS.0000000000005587.

Genz A, Bretz F (2009). Computation of Multivariate Normal and t Probabilities, series Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Geyh S, Fellinghauer BA, Kirchberger I, Post MW. Cross-cultural validity of four quality of life scales in persons with spinal cord injury. *Health Qual Life Outcomes* 2010; 8: 94. Https://doi: 10.1186/1477-7525-8-94.

Gibbons RD, Bock RD, Hedeker D, et al. Full-Information Item Bifactor Analysis of Graded Response Data. Applied Psychological Measurement 2007; 31(1): 4-19.

Giesinger JM, Petersen M, Groenvold M, et al. European Organisation for Research and Treatment of Cancer Quality of Life Group (EORTC-QLG). Cross-cultural development of an item list for computer-adaptive testing of fatigue in oncological patients. *Health Qual Life Outcomes* 2011; 9: 19. doi: 10.1186/1477-7525-9-19.

Giordano A, Pucci E, Naldi P, et al. Responsiveness of patient- reported outcome measures in multiple sclerosis relapses: the REMS study. *J Neurol Neurosurg Psychiatry* 2009; 80: 1023-8.

Giordano A, Testa S, Bassi M, et al. Assessing measurement invariance of MSQOL-54 across Italian and English versions. *Qual Life Res* 2020; 29(3): 783-91.

Gold SM, Heesen C, Schulz H, et al. Disease-specific quality of life instruments in multiple sclerosis: validation of the Hamburg Quality of Life Questionnaire in multiple sclerosis (HAQUAMS). *Mult Scler* 2001; 7: 119-30.

Gold SM, Schulz H, Mönch A, Schulz KH, Heesen C. Cognitive impairment in multiple sclerosis does not affect reliability and validity of self-report health measures. *Mult Scler* 2003; 9(4): 404-10.

Gough Ir, Furnival Cm, Schilder L, Grove W. Assessment of the quality of life of patients with advanced cancer. *Eur J Cancer Clin Oncol* 1983; 19: 1161-5.

Gustafsson J, Balke G. General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research* 1993; 28: 407–34.

Hadgkiss EJ, Jelinek GA, Weiland TT, Pereira NG, Marck CH, van derMeer DM. Methodology of an International Study of People with Multiple Sclerosis Recruited through Web 2.0 Platforms: Demographics, Lifestyle, and Disease Characteristics. *Neurology Research International* 2013; 580-596.

Hakim EA, Bakheit AM, Bryant TN, et al. The social impact of multiple sclerosis: a study of 305 patients and their relatives. *Disabil Rehabil* 2000; 22: 288–93.

Haley SM, Gandek B, Siebens H, et al. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: II. Participation outcomes. *Arch Phys Med Rehabil* 2008; 89: 275-83.

Harrison C, Loe BS, Lis P, Sidey-Gibbons C. Maximizing the Potential of Patient-Reported Assessments by Using the Open-Source Concerto Platform With Computerized Adaptive Testing and Machine Learning. *J Med Internet Res* 2020; 22(10): e20950. doi: 10.2196/20950.

Harwell MR, Janosky JE. An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. Applied Psychological Measurement 1991; 15(3): 279–291.

Hays RD, Revicki D, Coyne KS. Application of structural equation modeling to health outcomes research. *Evaluation Health Professions* 2005; 28: 295–309.

Hays RD, Stewart AL. The Structure of Self-Reported Health in Chronic Disease Patients. *Psychological Assessment: A Journal of Consulting and Clinical Psychology* 1990; 2: 22-30.

Healy BC, Zurawski J, Gonzalez CT, Chitnis T, Weiner HL, Glanz BI. Assessment of computer adaptive testing version of the Neuro-QOL for people with multiple sclerosis. *Mult Scler* 2019; 25(13): 1791-9.

Heesen C, Solari A, Giordano A, Kasper J, Köpke S. Decisions on multiple sclerosis immunotherapy: new treatment complexities urge patient engagement. *J Neurol Sci* 2011; 306: 192–7.

Hickey AM, Bury G, O'Boyle CA, Bradley F, O'Kelly FD, Shannon W. A new short form individual quality of life measure (SEIQoL-DW): application in a cohort of individuals with HIV/AIDS. *BMJ* 1996; 313(7048): 29-33.

Hobart JC, Lamping DL, Fitzpatrick R, Riazi A, Thompson A. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient- based outcome measure. *Brain* 2001; 124: 962-73.

Hohol MJ, Hohol MJ, Orav EJ, Weiner HL. Disease steps in multiple sclerosis: a simple approach to evaluate disease progression. *Neurol* 1995; 45: 251-5.

Holzinger KJ, Swineford F. The Bi-factor method. *Psychometrika* 1937; 2: 41–54.

Hu L, Bentler PM (1995). Evaluating model fit. In: R.H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp.76-99). London, UK: Sage.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 1999; 6: 1-55.

Hu LT, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods* 1998; 3: 424–53.

Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. *Soc Sci Med* 1981; 15: 221-9.

Hunt SM, McKenna SP. The QLDS: a scale for the measurement of quality of life in depression. *Health Policy* 1993; 22: 307-19.

Iacus SM, King G, Porro G. cem: Software for Coarsened Exact Matching. *Journal Statistical Software* 2009; 30 (9). https://doi.org/10.18637/jss.v030.i09.

Idiman E, Uzunel F, Ozakbas S, et al. Cross-cultural adaptation and validation of multiple sclerosis quality of life questionnaire (MSQOL-54) in a Turkish multiple sclerosis sample. *J Neurol Sci* 2006; 240: 77-80.

Jelinek GA, De Livera AM, Marck CH, et al. (2016). Lifestyle, medication and socio-demographic determinants of mental and physical health-related quality of life in people with multiple sclerosis. *BMC Neurology* 2016; 16: 235.

Jette AM, Haley SM, Tao W, et al. Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. *Phys Ther* 2007; 87: 385-98.

Joreskog KG. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 1969; 34: 183-202.

Katz S, Downs T, Cash H, Grotz R. Progress in development of the index of ADL. *Gerontologist* 1976; 10: 20–30.

Khurana V, Sharma H, Afroz N, Callan A, Medin J. Patient-reported outcomes in multiple sclerosis: a systematic comparison of available measures. *Eur J Neurol* 2017; 24: 1099-107.

Koch-Henriksen N, Rasmussen S, Stenager E, Madsen M: The Danish Multiple Sclerosis Registry. History, data collection and validity. *Dan Med Bull* 2001; 48(2): 91–4.

Kopec JA, Badii M, McKenna M, et al. Computerized adaptive testing in back pain: validation of the CAT-5D-QOL. *Spine* 2008; 33: 1384-90.

Köpke S, Kasper J, Mühlhauser I, Nübling M, Heesen C. Patient education program to enhance decision autonomy in multiple sclerosis relapse management: a randomized-controlled trial. *Mult Scler* 2009; 15(1): 96–104.

Kosinski M, Bjorner JB, Ware JE, Sullivan E, Straus WL. An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. *J Clin Epidemiol* 2006; 59: 715-23.

Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale. *Neurology* 1983; 33: 1444-52.

Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Qual Life Res* 2003; 12(5): 485-501.

Landrø NI, Celius EG, Sletvold H. Depressive symptoms account for deficient information processing speed but not for impaired working memory in early phase multiple sclerosis (MS). *J Neurol Sci* 2004; 217: 211–16.

Likert R, Roslow S, Murphy G. A simple and reliable method of scoring the Thurstone attitude scales. *J Soc Psychol* 1934; 5: 228–238.

Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932; 140: 5–55.

Lintern TC, Beaumont G, Kenealy PM, Murrell RC. Quality of life (QoL) in severely disabled multiple sclerosis patients: Comparison of three QoL measures using multidimensional scaling. *Qual Life Res* 2001; 10: 371-8.

Loe BS, Stillwell D, Gibbons C. Computerized Adaptive Testing Provides Reliable and Efficient Depression Measurement Using the CES-D Scale. *J Med Internet Res* 2017; 19(9): e302. doi: 10.2196/jmir.7453.

Lord F, Novick M. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley; 1968.

Luo D, Petrill SA, Thompson LA. An exploration of genetic g: Hierarchical factor analysis of cognitive data from the Western Reserve Twin Project. *Intelligence* 1994*;* 18: 335–47.

Mahoney FI, Barthel DW. Functional evaluation: the Barthel index. *Md State Med J* 1965; 14: 61-5.

Marrie RA, Cutter G, Tyry T, et al. Changes in the ascertainment of multiple sclerosis. *Neurology* 2005; 65: 1066–70.

Marrie RA, Cutter G, Tyry T, Vollmer T, Campagnolo D. Does multiple sclerosis-associated disability differ between races? *Neurology* 2006; 66(8): 1235–1240.

Martin M, Kosinski M, Bjorner JB, Ware JE, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. *Qual Life Res* 2007; 16: 647-60.

Massacesi L, Tramacere I, Amoroso S, et al. Azathioprine versus beta interferons for relapsing-remitting multiple sclerosis: a multicentre randomized non-inferiority trial. *PLoS ONE* 2014; 9: e113371.

Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–74.

McBride JR (1997). Research antecedents of applied adaptive testing. In: Sands WA, Waters BK, and McBride JR (eds.), *Computer Adaptive Testing: from inquiry to operation*. Washington, American Psychological Association.

McDonald RP, Marsh HW. Choosing a multivariate model: Noncentrality and goodness of fit. Psychological Bulletin 1990 107: 247–55.

Mendes MF, Balsimelli S, Stangehaus G, Tilbery CP. Validation of the functional assessment of multiple sclerosis quality of life instrument in a Portuguese language. *Arq Neuropsiquiatr* 2004; 62: 108-13.

Michel P, Baumstarck K, Ghattas B, et al. Multidimensional Computerized Adaptive Short-Form Quality of Life Questionnaire Developed and Validated for Multiple Sclerosis: The MusiQoL-MCAT. *Medicine* 2016; 95(14): e3068. doi: 10.1097/MD.0000000000003068.

Michel P, Baumstarck K, Lancon C, et al. Modernizing quality of life assessment: development of a multidimensional computerized adaptive questionnaire for patients with schizophrenia. *Qual Life Res* 2018; 27(4): 1041-54. doi: 10.1007/s11136-017-1553-1.

Miller DM, Allen R. Quality of life in multiple sclerosis: determinants, measurement, and use in clinical practice. *Current Neurology and Neuroscience Reports* 2010; 10: 397–406.

Miller DM, Bethoux F, Victorson D, et al. Validating Neuro-QoL short forms and targeted scales with people who have multiple sclerosis. *Mult Scler* 2016; 22(6): 830-41.

Millsap RE, Yun-Tein J. Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research* 2004; 39(3): 479–515.

Millsap RE, Yun-Tein J. Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research* 2004; 39(3): 479–515.

Minden SL, Frankel D, Hadden L, et al.: The Sonya Slifka Longitudinal Multiple Sclerosis Study: methods and sample characteristics. *Mult Scler* 2006; 12(1): 24–38.

Mitchell AJ, Benito-León J, González JM, Rivera-Navarro J. Quality of life and its assessment in multiple sclerosis: integrating physical and psychological components of wellbeing. *Lancet Neurology* 2005; 4: 556-66.

Mokkink LB, Knol DL, Uitdehaag BM. Factor structure of Guy's Neurological Disability Scale in a sample of Dutch patients with multiple sclerosis. *Mult Scler 2*011; 17(12): 1498-503.

Montalban X, Gold R, Thompson AJ, et al. ECTRIMS/EAN Guideline on the pharmacological treatment of people with multiple sclerosis. *Eur J Neurol* 2018; 25: 215–37.

Morales-Gonzalez JM, Benito-León J, Rivera-Navarro J, Mitchell AJ. A systematic approach to analyze health-related quality of life in multiple sclerosis: the GEDMA study. *Mult Scler* 2004; 10: 47–54.

Motl RW, McAuley E, Mullen S. Longitudinal measurement invariance of the Multiple Sclerosis Walking Scale-12. *J Neurol Sci* 2011; 305: 75-9.

Motl RW, McAuley E, Suh Y. Validity, invariance and responsiveness of a self-report measure of functional limitations and disability in multiple sclerosis. *Disab Rehabilitation* 2010; 32: 1260–71.

Motl RW, Mullen S, McAuley E. Multi-group measurement invariance of the multiple sclerosis walking scale-12? Neurological Research 2012; 34(2): 149-52.

Multiple Sclerosis Quality of Life (MSQOL)-54 Instrument. Available at: https://www.nationalmssociety.org/NationalMSSociety/media/MSNationalFiles/Brochures/MSQOL54_995.pdf (accessed 15 March 2021).

Muraki E (1997). A generalized partial credit model. In: van der Linden W & Hambleton RK (eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer.

Muraki E. A generalized partial credit model: Application of the EM algorithm. Applied Psychological Measurement 1992; 16: 159–76.

Muthén LK, Muthén BO (1998-2011). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

National Collaborating Centre for Chronic Conditions (UK). *Multiple Sclerosis. National clinical guideline for diagnosis and management in primary and secondary care 2004.* NICE Clinical Guidelines, No. 8.

Nortvedt MV, Riise T. The use of quality of life measures in multiple sclerosis research. *Mult Scler* 2003; 9: 63–72.

Nunnally J. *Psychometric Theory*. 2nd ed. New York, NY: McGraw-Hill; 1978.

Patrick D, Deyo R. Generic and disease-specific measures in assessing health status and quality of life. *Med Care* 1989; 27: S217–S232.

Patrick DL, Erickson P. *Health Status and Health Policy: quality of life in healthcare evaluation and resource allocation*. New York, Oxford University Press. 1993.

Peipert JD, Cella D. Bifactor analysis confirmation of the factorial structure of the Functional Assessment of Cancer Therapy–General (FACT-G). *Psycho-Oncology* 2019*;* 28: 1149–52.

Pekmezovic T, Kisic Tepavcevic D, Kostic J, Drulovic J. Validation and cross-cultural adaptation of the disease-specific questionnaire MSQOL-54 in Serbian multiple sclerosis patients sample. *Qual Life Research 2007;* 16: 1383–1387.

Petersen MA, Aaronson NK, Conroy T, et al. European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group. International validation of the EORTC CAT Core: a new adaptive instrument for measuring core quality of life domains in cancer. *Qual Life Res* 2020; 29(5): 1405-17.

Petersen MA, Groenvold M, Aaronson N, Fayers P, Sprangers M, Bjorner JB; European Organisation for Research and Treatment of Cancer Quality of Life Group. Multidimensional computerized adaptive testing of the EORTC QLQ-C30: basic developments and evaluations. *Qual Life Res* 2006; 15(3): 315-29.

Pfennings LEMA, Cohen L, Van der Ploeg HM. Assessing the quality of life in patients with multiple sclerosis. In *Multiple sclerosis: clinical changes and controversies* Edited by: Thompson AJ, Polman C, Hohlfeld R. London: Martin Dunitz; 1997.

Polman C, Reingold S, Edan G, et al. (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the ”McDonald Criteria“. *Annals Neurology*, 58, 840–6.

Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals Neurology* 2011; 69: 292–302.

Priestman TJ and Baum M. Evaluation of quality of life in patients receiving treatments for advanced breast cancer. *Lancet* 1976; 1: 899-901.

Prunty MC, Sharpe L, Butow P, Fulcher G. The motherhood choice: a decision aid for women with multiple sclerosis. *Pat Educ Counsel* 2008; 71(1): 108–15.

Putnick DL, Bornstein MH. Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review* 2016; 41: 71-90.

Qualimetric. https://www.qualitymetric.com/?utm_campaign=eu_shortform&utm_term=sf-36%20manual&gclid=CjwKCAiA8ov_BRAoEiwAOZogwZ7eLzcCdjQS50Mh1IHzafTaEoPnTb8D6XqRlQ_rvHDFSkAS87BxUxoCSk0QAvD_BwE)

R Development Team. (n.d.). The R Project for Statistical Computing.

Raftery AE. Bayesian model selection in social research. *Sociological Methodology* 1995; 25: 111–63.

Ramsaransing G, De Keyser J. Benign course in multiple sclerosis: a review. *Acta Neurol Scandinavica* 2006; 113: 359–69.

Rebollo P, Castejón I, Cuervo J, et al.; Spanish CAT-Health Research Group. Validation of a computer-adaptive test to evaluate generic health-related quality of life. *Health Qual Life Outcomes* 2010; 8: 147. doi: 10.1186/1477-7525-8-147.

Reckase MD, McKinley RL. The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement* 1991; 15: 361-73.

Reckase MD. The difficulty of test items that measure more than one ability. *Applied Psychological Measurement* 1985; 9(4): 401–412.

Reeve BB, Hays RD, Bjorner JB, et al. PROMIS Cooperative Group. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 2007; 45: S22-31.

Reich DS, Lucchinetti CF, Calabresi PA. Multiple sclerosis. *New Engl J Med* 2018; 378: 169–80.

Reise SP, Moore TM, Haviland MG. Bifactor Models and Rotations: Exploring the Extent to which Multidimensional Data Yield Univocal Scale Scores. *J Pers Assess* 2010; 92: 544–59.

Reise SP. The Rediscovery of Bifactor Measurement Models. *Multivariate Behav Res* 2012; 47(5): 667–96.

Revicki D, Cella D. Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Qual Life Res* 1997; 6: 595–600.

Rivera-Navarro J, Morales-González J.M, Benito-León J, Madrid Demyelinating Diseases Group (GEDMA). Informal caregiving in multiple sclerosis patients: data from the Madrid demyelinating disease group study. *Disabil Rehabil* 2003; 25: 1057–64.

Rodriguez A, Reise SP, Haviland MG. Evaluating Bifactor Models: Calculating and Interpreting Statistical Indices. *Psychological Methods* 2016; 21: 137–50.

Rosato R, Testa S, Bertolotto A, et al. Development of a short version of MSQOL-54 using factor analysis and item response theory. *PLoS One* 2016*;* 11: e0153466.

Rosato R, Testa S, Bertolotto A, et al. Prospective validation of the abbreviated, electronic version of the MSQOL-54. *Mul Scler* 2018; 25(6): 856-66.

Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients and doctors perceptions and assessments of disability in multiple sclerosis. *BMJ* 1997; 314: 1580-3.

Rotstein Z, Barak Y, Noy S, Achiron A. Quality of life in multiple sclerosis: development and validation of the "RAYS" scale and comparison with the SF-36. *Int J Qual Health Care* 2000; 12: 511-7.

Rust J, Golombok (eds). Modern psychometric. *The science of psychological assessment*. Routledge, 2009.

Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM. A new approach to the measurement of quality of life. The Patient-Generated Index. *Med Care* 1994; 32(11): 1109-26.

Samejima F (1997). Graded response model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometric Monography* 1969; 34.

Santos D, Abad FJ, Miret M, et al. Measurement invariance of the WHOQOL-AGE questionnaire across three European countries. *Qual Life Res* 2017; 27: 1015–25.

Sass DA, Schmitt TA, Marsh HW. Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators. *Structural Equation Modeling* 2014; 21(2): 167–80.

Scalfari A, Neuhaus A, Daumer M, et al. Onset of secondary progressive phase and long-term evolution of multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2014; 85: 67–75.

Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods* 2002; 7(2): 147-77.

Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978; 6: 461–4.

Segall DO. Multidimensional adaptive testing. *Psychometrika* 1996; 61: 331-54.

Seo DG, Weiss DJ. Best design for multidimensional computerized adaptive testing with the bifactor model. *Educ Psychol Meas* 2015; 75(6): 954-78. doi: 10.1177/0013164415575147.

Sherbourne C. Social functioning: Sexual problems measures. In: Stewart, A.L., Ware, J.E., (Eds), *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach* (pp. 194-204). North Carolina: Duke University Press, 1992.

Shirani A, Zhao Y, Karim ME, et al. Association between use of interferon beta and progression of disability in patients with relapsing- remitting multiple sclerosis. *JAMA* 2012; 308: 247–56.

Simeoni M, Auquier P, Fernandez O, et al. Validation of the Multiple Sclerosis International Quality of Life questionnaire. *Mult Scler* 2008; 14: 219–30.

Smets EM, Garssen B, Bonke B, De Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *J Psychosom Res* 1995; 39(3): 315-25.

Smits N, Paap MCS, Böhnke JR. Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Qual Life Res* 2018; 27(4): 1055-63.

Solari A, Filippini G, Mendozzi L, et al. Validation of Italian multiple sclerosis quality of life 54 questionnaire. *J Neurol Neurosurg Psychiatry* 1999; 67:158-62.

Solari A, Motta A, Mendozzi L, et al. Computer-aided retraining of memory and attention in people with multiple sclerosis: a randomized, double-blind controlled trial. *J Neurol Sci* 2004; 222: 99-104.

Solari A. Role of health-related quality of life measures in the routine care of people with multiple sclerosis. *Health Qual Life Outcomes* 2005; 3: 16. doi: 10.1186/1477-7525-3-16.

Spearman C. General intelligence: objectively determined and measured. *American Journal of Psychology* 1904; 115: 201-92.

Sprangers MAG, de Regt EB, Andries F, et al. Which chronic conditions are associated with better or poorer quality of life? *J Clin Epidemiol* 2000; 53: 895–907.

Steiger JH (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research* 1980; 25: 173–80.

Steiger JH, Lind JC (1980). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger JH. A note on multiple sample extensions of the RMSEA fit index. Structural Equation Modeling 1998; 5: 411–19.

Stewart A, Greenfield S, Hays R, et al. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. *J Am Med Assoc* 1989; 262: 907–13.

Stone M. Comments on Model Selection Criteria of Akaike and Schwarz Author (s): M. Stone Source: Journal of the Royal Statistical Society. Series B (Methodological), Vol. 41, No. 2. Blackwell Publishing for the Royal Statistical Society Sta, R. Stat. Soc, 2009: 276–8.

Sunderland M, Afzali MH, Batterham PJ, et al. Comparing Scores From Full Length, Short Form, and Adaptive Tests of the Social Interaction Anxiety and Social Phobia Scales. *Assessment* 2020; 27(3): 518-532. doi: 10.1177/1073191119832657.

Thissen D, Steinberg L. A taxonomy of item response models. Psychometrika 1986; 51(4): 567–77.

Tramacere I, Del Giovane C, Salanti G, D'Amico R, Filippini G. Immunomodulators and immunosuppressants for relapsing-remitting multiple sclerosis: a network meta- analysis. *Cochrane Database Syst Rev* 2015; 9. DOI: 10.1002/14651858.CD011381.

Trojano M, Bergamaschi R, Amato MP, et al. Italian Multiple Sclerosis Register Centers Group. The Italian multiple sclerosis register. *Neurol Sci* 2019; 40(1): 155-65.

United States Department of Health and Human Services Food and Drug Administration (FDA). Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. https://www.fda.gov/media/77832/download.

Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008; 17 (2): 179–93.

van der Linden WJ, Glas CAW (Eds) (2010). Elements of adaptive testing. New York: Springer.

Vernay D, Gerbaud L, Biolay S, et al. Quality of life and multiple sclerosis: validation of the French version of the self-questionnaire (SEP-59). *Rev Neur* 2000; 156: 247-63.

Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health- related quality of life measure for multiple sclerosis. *Qual Life Res* 1995; 4: 187-206.

Wainer H, Dorans NJ (eds). *Computerized Adaptive Testing*: A Primer (2nd Edition). Routledge, 2000.

Walter OB, Becker J, Bjorner JB, Fliege H, Klapp BF, Rose M. Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Qual Life Res* 2007; 16: 143-55.

Ware JE, Kosinski M, Bjorner JB, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual Life Res* 2003; 12:935-52.

Ware JE, Snow KK, Kosinski M, Gandek B. (1993). *SF-36 Health survey manual and interpretation guide*. Boston, MA: The Health Institute.

World Health Organization. Constitution of the World Health Organization. Geneva, WHO Basic Documents, 1948.

Yamamoto T, Ogata K, Katagishi M, et al. Validation of the Japanese-translated version Multiple Sclerosis Quality of Life-54 instrument. *Rinsho Shinkeigaku* 2004; 44: 417-21.

Yao L. Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *J Educ Meas* 2014; 51: 18–38.

Yen WM. A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika* 1987; 52(2): 275–91.

Yen WM. Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl Psychol Measure* 1984; 8: 125- 45.

Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983; 67(6): 361-70.

APPENDIX

Descriptive statistics of the MSQOL-54 questionnaire.

Item	Mean (SD)	Min-max	Skewness	Kurtosis	Missing (%)
1	3.05 (1.02)	1-5	-0.07	2.50	2.1
2	3.01 (1.00)	1-5	-0.30	2.86	2.1
3	1.66 (0.78)	1-3	0.66	1.95	0.8
4	2.20 (0.79)	1-3	-0.37	1.68	0.3
5	2.30 (0.77)	1-3	-0.56	1.90	0.4
6	2.08 (0.83)	1-3	-0.16	1.47	0.4
7	2.41 (0.75)	1-3	-0.84	2.26	0.5
8	2.30 (0.78)	1-3	-0.55	1.86	0.3
9	2.03 (0.88)	1-3	-0.07	1.30	0.3
10	2.24 (0.86)	1-3	-0.48	1.54	0.4
11	2.45 (0.77)	1-3	-0.96	2.36	0.4
12	2.66 (0.61)	1-3	-1.60	4.38	0.4
13	1.58 (0.49)	1-2	-0.34	1.12	0.7
14	1.43 (0.49)	1-2	0.28	1.07	0.6
15	1.48 (0.50)	1-2	0.07	1.00	0.5
16	1.49 (0.50)	1-2	0.04	1.00	1.0
17	1.70 (0.46)	1-2	-0.90	1.82	0.7
18	1.61 (0.49)	1-2	-0.45	1.20	0.7
19	1.65 (0.48)	1-2	-0.61	1.37	0.8
20	2.32 (1.16)	1-5	0.51	2.32	0.2
21	2.63 (1.40)	1-6	0.34	1.93	0.2
22	2.01 (1.14)	1-5	0.84	2.67	0.3
23	3.76 (1.31)	1-6	-0.15	2.23	0.2
24	4.51 (1.29)	1-6	-0.78	2.99	0.3
25	4.85 (1.23)	1-6	-1.02	3.44	0.5
26	3.48 (1.23)	1-6	-0.02	2.23	0.5
27	4.04 (1.35)	1-6	-0.34	2.30	0.5

28	4.52 (1.20)	1-6	-0.80	3.25	0.5
29	3.80 (1.39)	1-6	-0.25	2.19	0.7
30	3.14 (1.26)	1-6	0.18	2.14	0.5
31	3.21 (1.34)	1-6	-0.07	2.07	0.8
32	3.73 (1.48)	1-6	-0.16	1.94	0.5
33	3.57 (1.14)	1-6	-0.34	2.31	1.6
34	3.66 (1.28)	1-5	-0.61	2.24	0.3
35	3.03 (1.33)	1-5	0.07	1.75	0.6
36	2.98 (1.14)	1-5	0.17	2.39	0.4
37	3.24 (1.33)	1-5	-0.04	1.63	0.5
38	4.17 (1.38)	1-6	-0.58	2.58	0.2
39	4.12 (1.51)	1-6	-0.51	2.24	0.4
40	3.93 (1.45)	1-6	-0.43	2.29	0.2
41	4.33 (1.53)	1-6	-0.07	2.46	0.2
42	4.29 (1.43)	1-6	-0.59	2.48	0.2
43	4.34 (1.44)	1-6	-0.64	2.50	0.4
44	4.28 (1.43)	1-6	-0.64	2.51	0.3
45	4.73 (1.46)	1-6	-1.05	3.08	0.2
46	2.01 (1.13)	1-4	0.64	1.94	6.4
47	1.89 (1.09)	1-4	0.83	2.23	8.5
48	1.93 (1.12)	1-4	0.80	2.15	7.9
49	1.71 (1.04)	1-4	1.17	2.93	8.8
50	2.74 (1.33)	1-5	0.34	2.00	8.0
51	1.90 (1.13)	1-5	1.12	3.28	0.3
52	1.98 (1.10)	1-5	0.91	2.84	0.2
53	6.69 (2.03)	1-10	-0.74	3.29	0.7
54	4.67 (1.34)	1-7	-0.56	3.03	1.2