The minimal clinically important difference determined using item response theory models: an attempt to solve the issue of the association with baseline score.

Alexandra Rouquette, Myriam Blanchin, Véronique Sébille, Francis Guillemin, Sylvana Côté, Bruno Falissard, Jean-Benoit Hardouin

The Minimal Clinically Important Difference determined using Item Response Theory Models: an attempt to solve the issue of the association with baseline score

Alexandra Rouquette[1,2,3,4,5*], Myriam Blanchin[4], Véronique Sébille[4], Francis Guillemin[6], Sylvana M Côté[3,7], Bruno Falissard[1,2,8] and Jean-Benoit Hardouin[4]

[1] Inserm, U669, Paris, France

[2] UMR-S0669, Université Paris-Sud and Université Paris Descartes, Paris, France ;

[3] Research Unit on Children's Psychosocial Maladjustment, University of Montreal, Montreal, Quebec, Canada

[4] EA 4275 « Biostatistics, Clinical Research and Subjective Measures in Health Sciences », University of Nantes, Nantes, France

[5] Department of Public Health, University of Angers, Angers, France

[6] EA 4360 APEMAC, INSERM CIC-EC CIE6 and CHU Nancy, Université de Lorraine and Université Paris Descartes, Nancy, France

[7] Department of preventive and social medicine, University of Montreal, Montreal, Québec, Canada

[8] Department of Public Health, Paul Brousse Hospital, Villejuif, France

*Corresponding Author

ABSTRACT

Objective: Determining the Minimal Clinically Important Difference (MCID) of questionnaires on an interval scale, the Trait Level (TL) scale, using Item Response Theory (IRT) models could overcome its association with baseline severity. The aim of this study was to compare the Sensitivity (Se), Specificity (Sp) and Predictive Values (PV) of the MCID determined on the Score scale (MCID-Sc) or on the TL-scale (MCID-TL).

Study design and setting: The MCID-Sc and MCID-TL of the MOS-SF36 general health subscale were determined for deterioration and for improvement on a cohort of 1170 patients using an anchor-based method and a partial credit model. The Se, Sp and PV were calculated using the global rating of change (the anchor) as the gold standard test.

Results: The MCID-Sc magnitude was smaller for improvement (1.58 points) than for deterioration (-7.91 points). The Se, Sp and PV were similar for MCID-Sc and MCID-TL in both cases. However, if the MCID was defined on the score scale as a function of a range of baseline scores, its Se, Sp and PV were consistently higher.

Conclusion: This study reinforces the recommendations concerning the use of a MCID-Sc defined as a function of a range of baseline scores.

KEYWORDS: Minimal clinically important difference; Questionnaires; Sensitivity and specificity; Item response theory; Rasch models; Patient-reported outcomes

RUNNING TITLE: MCID determination using IRT models

WORD COUNT: 3926 (+197 for the abstract)

TEXT BOX

What is new?

- The Minimal Clinically Important Difference (MCID) defined as a function of a range of baseline scores leads to a better classification of individuals having experienced “at least a minimally important change” versus “no change" over time than the MCID defined without considering the baseline severity
- Determining the MCID using Item Response Theory (IRT) models does not greatly enhance its Sensitivity (Se), Specificity (Sp) and Predictive Values (PV) compared with its determination on the score scale
- The lack of interval scale properties of the score is not fully responsible for the MCID dependence on baseline severity

1 Introduction

Multi-item questionnaires are increasingly used in longitudinal studies to measure perceived health status and to assess its changes over time. Indeed, clinicians and policy makers are more and more interested in integrating patient's perspective and experience of disease and illness in the evaluation of treatments, interventions or public health policies [1-5]. However, a major limitation to the use of these measurement instruments in clinical research or epidemiological studies is their interpretability [6-17]. For instance, what is the meaning of a two-point reduction over a six month period when anxiety is assessed with a 20 points scale? Is it a trivial or a meaningful difference? The Minimal Clinically Important Difference (MCID) is a concept defined to help with the interpretation of observed differences obtained in longitudinal studies using questionnaires [18].

The best method for determining the MCID of a questionnaire is still under debate, however, anchor-based methods are recommended by numerous authors as they compare observed score differences to external criteria that have clinical relevance [7, 11-12, 14, 19]. These criteria can be indicators of clinical response or of illness evolution but the most used are patient-based Global Ratings of Change (GRC) since they provide a simple measure of the significance of change from the individual perspective [3, 13-14, 19-20]. In practice, multiple anchors are more and more often used in the same study [21-23].

Several issues are, nevertheless, still complicating the MCID determination, especially its variations among populations, estimations approaches, etc., and raise questions about the existence of a unique questionnaire-specific MCID [11, 24-27]. One of these issues concerns the influence of the subjects' baseline score (BS) on the MCID value calculated using the anchor-based methods, as it has been shown in various studies [9, 27-36]. For that matter, various authors have recommended to define the MCID as a function of a range of

BS rather than one single MCID [12, 14, 19, 30, 35]. Thus, to be able to conclude on the meaningfulness of someone's change, different MCID values should be considered depending on the subject's BS.

Several origins to this phenomenon have been mentioned [12]. The first one can be explained in psychophysical terms and is in relation with the subjective feature of the MCID concept: the subject's perception of a clinically meaningful change can be different depending on his/her baseline severity [37]. The second one is related to the statistical nature of the MCID concept and is called regression to the mean that describes the statistical tendency of extreme scores to become less extreme at follow-up [19, 30]. At last, two other potential origins of the MCID association with baseline severity concern the score itself (possibly weighted sum of the items responses) used as a measure of the construct (i.e. pain, anxiety, etc.) evaluated by the questionnaire. One of these origins is due to the upper and lower bounds of this score which are responsible for the floor/ceiling effects: patients whose BS is close to the ends of the scale are not able to register a large change because such a change would exceed the span of the scale [12, 25, 36-37]. The other one concerns the scale level of the score which has not necessarily the interval scale properties. With an interval scale, units along the scale are equal to one another [4, 36, 38-39]. The present study focuses on the potential lack of interval scale properties of the score and its role in the MCID dependence to BS phenomenon. Indeed, if the score scale is not an interval scale, interpretation of score differences can vary depending on the different portions of the scale.

Models from the Item Response Theory (IRT) are convenient tools to analyze questionnaire data and express the results on an interval scale. In this theory, the construct measured by the questionnaire, called latent trait, is assessed by a quantitative variable with interval scale properties, the Trait Level (TL) [40]. Thus, if the questionnaire measures

anxiety (the latent trait) for example, an x-unit difference represents the same quantity whatever its location on the TL-scale (low, medium or high level of anxiety). If our hypothesis concerning the role of the interval scale properties in the MCID association with baseline severity is true, the MCID determination on the TL-scale using an IRT model, could therefore avoid this phenomenon. We could, thus, expect fewer misclassifications of individuals having experienced “at least a minimally important change” versus “no change" over time than with the MCID determined on the score scale.

The aim of our study is therefore to compare the Sensitivity (Se), Specificity (Sp), Positive and Negative Predictive Values (PPV and NPV) of the MCID determined on the Score scale (MCID-Sc) and the MCID determined on the TL-scale (MCID-TL) using an IRT model and an anchor-based method in which the external criteria is considered as the gold standard test.

2 Methods

2.1 Data source

Data came from a French multicenter longitudinal prospective SATISQOL study composed of 1709 hospitalised patients, enrolled between October 2008 and September 2010, younger than 75 years-old and attending surgery or medical intervention for a chronic illness of one of the following systems: cardiovascular, musculoskeletal, nephrology, urology, digestive, pulmonary or endocrine. To be included, patients needed to speak French, have sufficient cognitive function to complete a self-administered questionnaire and exhibit symptoms of their chronic illness for, at least, six months. They were excluded if they did not have a therapeutic intervention during their hospitalisation.

Demographic information (age, sex, diagnosis, etc.), self-reported satisfaction with care (French version of the Patient Judgements of Hospital Quality questionnaire [41-42])

and quality of life (French version of the Medical Outcomes Study Short Form-36 questionnaire - MOS-SF36 [43-44]) were obtained during hospitalisation. Six months later, patients were asked to fill in the MOS-SF36 questionnaire again during a scheduled medical consultation. The study was approved by the ethic committee of Lorraine, France, and all the patients gave their informed consent to participate.

2.2 Questionnaire

The MOS-SF36 is a generic 36-items questionnaire divided into eight subscales addressing physical, mental and social health, and one item assessing health transition. To ensure the construct's unidimensionality required by the IRT model used in this study, analyses were performed on the five items of the General Health (GH) subscale. Each of these items was rated on an ordinal scale with five categories. The score, ranging from zero (worst perceived general health) to 100, was computed as recommended by the MOS-SF36 user's guide [43]. Likewise, an individual mean imputation was performed if there were less than three missing responses in the GH-subscale as advocated.

The item assessing health transition at the six-month follow-up was chosen to be used as the GRC : "Compared to six months ago, how would you rate your health in general now?" Patients could choose between five responses: "Much better", "Somewhat better", "About the same", "Somewhat worse" and "Much worse".

Patients with three or more missing responses in the GH-subscale or who did not answer to the GRC at the six-month follow-up were excluded from the sample used for the analyses.

2.3 Analyses

Since it is well known that the amount and quality of change is likely to be different for improvement as compared to deterioration, the following analyses were performed in both circumstances [14-15, 19, 25, 28].

2.3.1 MCID-Sc determination

Changes in general health over the six-month interval were computed as the difference between baseline (T1) and six-month (T2) GH-subscale score. The MCID-Sc was computed as the mean score change from T1 to T2 in the subgroup of patients who answered “Somewhat better” to the GRC (SB group) for improvement and in the subgroup of patients who answered “Somewhat worse” (SW group) for deterioration. The dependence of score change to the BS was evaluated using Pearson’s correlation coefficients. Polychoric correlation coefficients were used to assess association between score change and responses to the GRC.

Since it is recommended by various authors, a MCID-Sc composed of several values according to a range of Baseline Scores was determined: the MCID-Sc_{BS} [12, 14, 19, 30, 35]. Concretely, the MCID-Sc_{BS} was defined as the three means of score change from T1 to T2 for patients having a BS in the first third ([0 – 33]), the second third (]33 – 67[) or the higher third of the scale ([67 – 100]), in the SB group for improvement and in the SW group for deterioration.

2.3.2 MCID-TL determination

2.3.2.1 Assumptions of IRT

IRT models rely on three fundamental assumptions: unidimensionality, local independence and monotonicity. The unidimensionality of the GH subscale was checked, at

each assessment time, using an eigenvalues analysis and the fit examination of a Confirmatory Factor Analysis (CFA) model with one factor. The Root Mean Square Error Approximation (RMSEA, acceptable fit if <0.06), the Comparative Fit Index (acceptable fit if >0.95), the Tucker Lewis Index (acceptable fit if >0.95) and the Standardized Root Mean Square Residual (acceptable fit if <0.08) were examined to evaluate the fit of the CFA model [45]. A non parametric IRT analysis was also performed by fitting a Monotonely Homogeneous Model of Mokken to our data. A good fit, evaluated by the Loevinger's H coefficients, indicates that the three IRT fundamental assumptions are [46]. Finally, the internal consistency of the GH subscale was checked by the computation of the Cronbach's alpha coefficient which was considered as acceptable if it was higher than 0.7 [47].

2.3.2.2 Fit of the Partial Credit Model (PCM) and item parameters estimation

A PCM, an IRT model for polytomous data (cf. supplementary material), was fitted on the data at T1 and T2 separately. A PCM was chosen because it is a model of the Rasch family which is very commonly used in the field of health related questionnaires [48]. This model defines MxJ item parameters with M the number of the response categories of the J items of the scale. In this model, the concept measured by the scale is represented by a random variable following a normal distribution. Fit tests, based on a chi-squared comparison, are known to be highly susceptible to large sample sizes. The PCM fit was, thus, adjusted for an expected sample of 400 individuals at both assessment times, which is a large enough sample to estimate the parameters of a PCM [49].

Measurement invariance of the GH subscale was checked using comparisons of the item parameters confidence intervals at both assessment times. As recommended if measurement invariance is met, averaged item parameters from across the two assessment

times were obtained by fitting a PCM on a data set made up of the T1 and T2 data sets [50-51].

2.3.2.3 MCID-TL determination

A latent regression IRT model was used to assess the TL mean variation over time within the SB/SW groups (cf. supplementary material). The MCID-TL was thus defined as the time effect on the TL-scale (TL mean change from T1 to T2) in the SB group for improvement and in the SW group for deterioration, respectively. To classify patients as having experienced “at least a minimally important change” or “no change", these MCID-TL had to be translated onto the score scale. A PCM was thus used to provide the relationship between the TL and the expected score at the GH subscale (cf. supplementary material). Using this translation tool, the score difference equivalent to the MCID-TL was determined for each BS varying from 0.5 to 99.5 by an increment of 0.5. Thus, knowing each patient’s BS, it was possible to determine if his/her score change over the six-month interval was larger than the MCID-TL or not. Due to the logistic form of the PCM, it was not possible to translate the MCID-TL for extreme BS (0 or 100); therefore it was approximated to the value obtained for the nearest BS (0.5 or 99.5 respectively).

2.3.3 Se, Sp, PPV and NPV computation

Each patient of the whole sample was classified as having experienced “at least a minimally important change” or “no change" over the six-month interval using the MCID-Sc, the MCID-Sc_{BS} and the MCID-TL classifications. Se, Sp, PPV and NPV were thus computed using the patient’s response at the GRC as the gold standard classification.

2.3.4 Software

Descriptive analysis, graphs, factor analysis and non parametric IRT analysis were performed using Stata®/MP 12.1 and the Microsoft® Office Excel 2007 spreadsheet program [52-53]. The items parameters and the PCM fit were estimated using RUMM® 2030 [54]. Finally, the SAS software 9.3 was used to estimate the MCID-TL values using the longitudinal form of the PCM with mixed effects [55].

3 Results

At baseline, 1709 patients (877 men – 56.1%, 686 women – 43.9%, missing information for 146 patients) were entered. The average age of the participants was 55.7 (Standard Deviation – SD=14.0) years with a range of 18 to 80 years. At six-month follow-up, the response rate was 89.4%, i.e. 1528 patients. Amongst them, 58 did not answer to the GRC at T2 and 300 had more than two missing responses to the GH-subscale at T1 or T2, leaving 1170 patients for the analysis. The average GH-subscale score was 52.1 (SD=22.4) at T1 and 51.7 (SD=23.3) at T2. In the **figure 1** is depicted a histogram of the BS which was lower than or equal to 33 for 269 (23.0%) patients and higher than or equal to 67 for 372 (31.8%) patients.

[Figure 1 near here]

3.1 MCID-Sc determination

The response to the GRC was “Much better” for 266 (22.7%) patients, “Somewhat better” for 360 (30.8%), “About the same” for 401 (34.3%), “Somewhat worse” for 112 (9.6%) and “Much worse” for 31 (2.6%). The MCID-Sc of the GH subscale was equal to 1.58 (Standard Error - SE=0.76) points for improvement and to -7.91 (SE=1.26) points for deterioration. To notice, the mean score change in the group of patients considered as stable (who rated their health as “About the same” compared with six months ago) was -

3.16 (SE=0.68). Polychoric correlation between score change and responses to the GRC was equal to -0.29.

Pearson's correlation between the score change and the BS was equal to -0.35 in the SB group and to -0.62 in the SW group. Box plots in **figure 2** show the variation of the score change over the six-month interval depending on the BS in the SB and SW groups. Globally, for improvement, the higher the BS, the smaller the score change. Conversely, for deterioration, the higher the BS, the larger the score change.

Means of score change specified in **figure 2** for each subgroup of patients defined by their BS were used to determine the MCID-Sc_{BS}. For instance, the MCID-Sc_{BS} for improvement was equal to 8.4 (SE=1.4) if the BS was included in [0 – 33] and to 2.5 (SE=1.0) if it was included in]33 – 67[in the SB group. If the BS was included in [67 – 100] in the SB group, the MCID-Sc_{BS} for improvement was set to zero since the mean score change was negative in this subgroup.

[Figure 2 near here]

3.2 MCID-TL determination

At both times, only one eigenvalue was higher than one and the ratio of the first to the second eigenvalue was higher than four. All the criteria indicated an acceptable fit for the one factor CFA model, except the RMSEA which was equal to 0.088 at T1 and to 0.102 at T2. However, all the Loevinger's H coefficients did not detect any violation of the fundamental IRT assumptions. Finally, a good internal consistency was found at both assessment times with a Cronbach's alpha coefficient equal to 0.81 at T1 and to 0.84 at T2.

The assumptions of a good PCM fit to the data were not rejected at 5% (p=0.19 at T1 and p=0.32 at T2). The measurement invariance of the GH-subscale was assumed since the confidence interval of the 20 item parameters estimated at T1 overlapped with their

confidence interval estimated at T2. The MCID-TL for improvement was estimated at 0.0839 (SE=0.0443) and at -0.4806 (SE=0.0833) for deterioration. It can be noted that the mean TL change in the group of patients considered as stable was equal to -0.1919 (SE=0.0426).

In the **figure 3** is depicted the relationship between the expected GH subscale score and the TL whose logistic shape is typical of the Rasch family models. Using this translation tool, it was possible to translate the MCID-TL on the score for each BS and represent it, as in the **figure 4** on the X-axis with the BS on the Y-axis. For example, a patient with a score of 20 on the GH subscale at baseline should have undergone a 1.5 points increase on his/her score at T2 to be classified as having experienced a minimal clinically important improvement using the MCID-TL whereas a patient with a BS equal to 80 should have undergone a 0.5 points increase.

[Figure 3 and figure 4 near here]

3.3 Se, Sp, PPV and NPV calculation

The Se, Sp and predictive values for the MCID-Sc, MCID-Sc_{BS} and the MCID-TL are shown for improvement and deterioration in **table 1**. All these values but one are lower than 80%.

[Table 1 near here]

4 Discussion

Our study was designed to evaluate the advantages of IRT models for the determination of the MCID of the MOS-SF36 questionnaire GH subscale in a sample of hospitalized patients suffering from a chronic disease and undergoing a therapeutic intervention. In our study, the use of IRT models does not improve the Se, Sp and predictive values of the MCID-TL compared to the MCID-Sc, except for deterioration where its Se and

predictive values seem slightly increased. For the MCID-Sc_{BS}, observed Se, Sp, and predictive values are consistently higher than for MCID-Sc or MCID-TL.

The overall lack of superiority of the MCID-TL compared to the MCID-Sc can be explained in considering **figures 1 and 3**. Indeed, in the **figure 3**, it can be seen that the relationship between the GH subscale score and the TL is quasi linear for a score ranged from 20 to 80 approximately. It means that, in this score range, the scale level of the GH subscale score nearly reaches the interval scale level. Moreover, in the study sample, 965 (82.5%) patients had a BS ranged in]20 – 80], as it can be seen in **figure 1**. It follows that few misclassifications of individuals having experienced “at least a minimally important change” versus “no change" over time, using the MCID-Sc, can be explained by the lack of interval scale properties of the score scale in our study. However, the magnitude of the MCID is another important factor to consider. Indeed, this magnitude is approximately five times larger for deterioration than for improvement. The slightly better MCID-TL’s performances in the case of deterioration suggested in our study could result from its magnitude since the lack of interval scale properties of the score could lead to more distortions in a large difference than in a small difference, i.e. the larger the quantity measured, the larger the discrepancy observed between its measures on the score scale or on the TL-scale.

The other important result of our study concerns the better results obtained with the MCID-Sc_{BS}. Further research should be done to disentangle the origins of this phenomenon and to determine if it could be explained by a different perception of change depending on the baseline severity. Indeed, for example, the MCID decrease with the increasing BS observed in the case of improvement could result from the ceiling effect as well as from the Regression To the Mean (RTM) phenomenon. In concrete terms, the ceiling effect is due to a lack of items able to measure a minimal clinically significant improvement for patients with

an already high score at baseline. The score change observed for these patients is, therefore, lower than the change which would have been observed if there had been no ceiling effect. Although this effect is smaller than on the score scale, the use of the LT is also subject to floor and ceiling effects and it might be another reason for the lack of superiority of the MCID-TL compared to the MCID-Sc [56]. The RTM phenomenon is responsible for a higher probability of negative change score for patients in the upper part of the BS distribution (statistical tendency of extreme scores to become less extreme at follow-up). In our study, the RTM could explain the negative mean change score (-4.8) observed in the subgroup of patients with a BS comprised in [67 – 100] in the SB group (i.e. a decreasing mean score on the GH subscale from T1 to T2 whereas patients rated their health in general on the GRC at T2 as better than at T1).

One of the most cited limits of the anchor-based method concerns the validity of the anchor [19, 27, 29, 57]. In our study, the weak values of the Se, Sp, ~~and~~ predictive values and correlations observed between score change from T1 to T2 and the GRC ~~observed~~ raise questions about the validity of the MOS-SF36 health transition item used as an anchor. In the MOS-SF36 questionnaire, the response to this item is not used to compute the score of the other eight dimensions assessed and, consequently, of the GH subscale. This item's face validity is obviously good to assess change on the construct supposed to be measured by the GH-subscale. However, the mean change in the subgroup of patients considered as stable (health in general rated as "About the same" compared with six months ago) was negative on the score scale (-3.16) as well as on the TL-scale (-0.19). These results raise different questions [27]: is the construct measured by the GH-subscale the same as the "health in general" referred to in the GRC? Has this GRC still the same meaning for the patients when assessing their health six month ahead (recall bias)? Finally, does response shift in one or

several items of the GH-subscale occur from T1 to T2? Further analysis should be done to clarify these issues. Another limit should be discussed concerning the heterogeneity of diseases in the cohort used in this study. The use of a more valid anchor and/or a more homogeneous clinically-defined cohort may have improved the values of the Se, Sp, PPV and NPV for each of the MCID values calculated with the three different methods but would unlikely have favoured one method over another.

To our knowledge, this work is the first one which uses IRT models to determine the MCID on the TL. These models are powerful tools that make the measurement of subjective phenomenon on an interval level scale possible. However, our study shows that, for the GH-subscale of the MOS-SF36 questionnaire, the ability of a single MCID value to classify individuals as having experienced “at least a minimally important change” versus “no change" over time, is not enhanced if the MCID is determined on the TL-scale compared with the MCID-Sc. Furthermore, the recommendations done by various authors concerning the use of several MCID values according to the baseline severity (MCID-Sc_{BS}) values are reinforced by our results [13, 15, 22, 30, 35]. Methods to determine the number of values for the MCID-Sc_{BS} which leads to the highest Se and Sp for a scale should be developed. The choice of this number should obviously be balanced with the logistical challenge of a large number of values in practice, especially with separate MCID values for improvement and deterioration.

Fundings

The French National Research Agency, under reference N-2010-PRSP-008-01, supported this study. The SATISQOL cohort project was supported by an IRESP (Institut de

recherche en santé publique) grant from Inserm, and a PHRC (Programme Hospitalier de Recherche Clinique) national grant from French Ministry of Health, France.

Conflict of interest: none declared

REFERENCES

1. Clancy CM and Eisenberg JM. (1998) Outcomes research: measuring the end results of health care. Science. 282:245-6.

2. Roger VL. (2011) Outcomes research and epidemiology: the synergy between public health and clinical practice. Circ Cardiovasc Qual Outcomes. 4:257-9.

3. US Department of Health and Human Services (USDHHS). Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. (2009); Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatory/Information/Guidances/UCM193282.pdf.

4. McHorney CA. (1997) Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med. 127:743-50.

5. Ellwood PM. (1988) Shattuck lecture--outcomes management. A technology of patient experience. N Engl J Med. 318:1549-56.

6. Guyatt GH and Cook DJ. (1994) Health status, quality of life, and the individual. JAMA. 272:630-1.

7. Liang MH. (2000) Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care. 38:II84-90.

8. Norman GR, Stratford P and Regehr G. (1997) Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 50:869-79.

9. Stucki G, Daltroy L, Katz JN, Johannesson M and Liang MH. (1996) Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol. 49:711-7.

10. Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, et al. (2001) Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. J Rheumatol. 28:400-5.

11. Beaton DE, Boers M and Wells GA. (2002) Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol. 14:109-14.

12. Copay AG, Subach BR, Glassman SD, Polly DW, Jr. and Schuler TC. (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J. 7:541-6.

13. Cook CE. (2008) Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. J Man Manip Ther. 16:E82-3.

14. Revicki D, Hays RD, Cella D and Sloan J. (2008) Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 61:102-9.

15. Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, et al. (2011) Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. J Clin Epidemiol. 64:487-96.

16. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, et al. (2010) Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 63:37-45.

17. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. (2011) A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol. 65:253-61.

18. Jaeschke R, Singer J and Guyatt GH. (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 10:407-15.

19. Crosby RD, Kolotkin RL and Williams GR. (2003) Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 56:395-407.

20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 60:34-42.

21. Yost KJ, Sorensen MV, Hahn EA, Glendenning GA, Gnanasakthy A and Cella D. (2005) Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. Value Health. 8:117-27.

22. Purcell A, Fleming J, Bennett S, Burmeister B and Haines T. (2010) Determining the minimal clinically important difference criteria for the Multidimensional Fatigue Inventory in a radiotherapy population. Support Care Cancer. 18:307-15.

23. Sloan JA. (2005) Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. COPD. 2:57-62.

24. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR and Aaronson NK. (2006) Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes. 4:70.

25. Hays RD and Woolley JM. (2000) The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics. 18:419-23.

26. Beaton DE. (2003) Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Med Care. 41:593-6.

27. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. (2010) Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 63:524-34.

28. Stratford PW, Binkley JM, Riddle DL and Guyatt GH. (1998) Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther. 78:1186-96.

29. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res. 16:131-42.

30. Crosby RD, Kolotkin RL and Williams GR. (2004) An integrated method to determine meaningful changes in health-related quality of life. J Clin Epidemiol. 57:1153-60.

31. Stratford PW, Binkley J, Solomon P, Finch E, Gill C and Moreland J. (1996) Defining the minimum level of detectable change for the Roland-Morris questionnaire. Phys Ther. 76:359-65; discussion 66-8.

32. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L and Grunnet-Nilsson N. (2006) Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord. 7:82.

33. Jensen MP, Chen C and Brugger AM. (2003) Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. J Pain. 4:407-14.

34. ten Klooster PM, Drossaers-Bakker KW, Taal E and van de Laar MA. (2006) Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. Pain. 121:151-7.

35. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. (2005) Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis. 64:29-33.

36. Bird SB and Dickson EW. (2001) Clinically significant changes in pain along the visual analog scale. Ann Emerg Med. 38:639-43.

37. Baker DW, Hays RD and Brook RH. (1997) Understanding changes in health status. Is the floor phenomenon merely the last step of the staircase? Med Care. 35:1-15.

38. Stevens SS. (1946) On the Theory of Scales of Measurement. Science. 103:677-80.

39. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J and Matchar D. (1999) Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics. 15:141-55.

40. Embretson SE and Reise SP. Item Response Theory for Psychologists: L. Erlbaum Associates; 2000.

41. Nguyen Thi PL, Briançon S, Empereur F and Guillemin F. (2002) Factors determining inpatient satisfaction with care. Soc Sci Med. 54:493-504.

42. Rubin HR, Ware JE, Jr., Nelson EC and Meterko M. (1990) The Patient Judgments of Hospital Quality (PJHQ) Questionnaire. Med Care. 28:S17-8.

43. Leplège A, Ecosse E, Coste J, Pouchot J and Perneger T. Le questionnaire MOS SF-36: Manuel de l'utilisateur et guide d'interprétation des scores: Estem; 2001.

44. Ware JE, Jr. and Sherbourne CD. (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 30:473-83.

45. Hu L and Bentler PM. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 6:1-55.

46. Sijtsma K and Molenaar IW. Introduction to Nonparametric Item Response Theory: SAGE Publications; 2002.

47. Cronbach LJ. (1951) Coefficient alpha and the internal structure of a test. Psychometrika 16:297–334.

48. Anthoine E, Moret L, Regnault A, Sébille V and Hardouin J-B. (Submitted) How PRO measures are psychometrically validated? A review of publications on primary validation.

49. Smith AB, Rush R, Fallowfield LJ, Velikova G and Sharpe M. (2008) Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol. 8:33.

50. Wright BD. (1996) Comparison requires stability. Rash Measurement Trans. 10:506.

51. Norquist JM, Fitzpatrick R, Dawson J and Jenkinson C. (2004) Comparing alternative Rasch-based methods vs raw scores in measuring change in health. Med Care. 42:I25-36.

52. Hardouin J-B, Bonnaud-Antignac A and Sébille V. (2011) Nonparametric item response theory using Stata. Stata Journal. 11:30-51.

53. StataCorp LP. Stata Statistical Software: Release 12.1. College Station, TX2012.

54. Andrich D, Sheridan BS and Luo G. Rumm2030: Rasch Unidimensional Measurement Models [computer software]. Perth, Western Australia: RUMM Laboratory; 2010.

55. SAS Institute Inc. Procedures Guides. Cary, NC: SAS Institute Inc; 2010.

56. Revicki DA and Cella DF. (1997) Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res. 6:595-600.

57. Kemmler G, Giesinger J and Holzner B. (2011) Clinically relevant, statistically significant, or both? Minimal important change in the individual subject revisited. J Clin Epidemiol. 64:1467-8.

Figure 1: Histogram of the general health (GH) subscale score at baseline

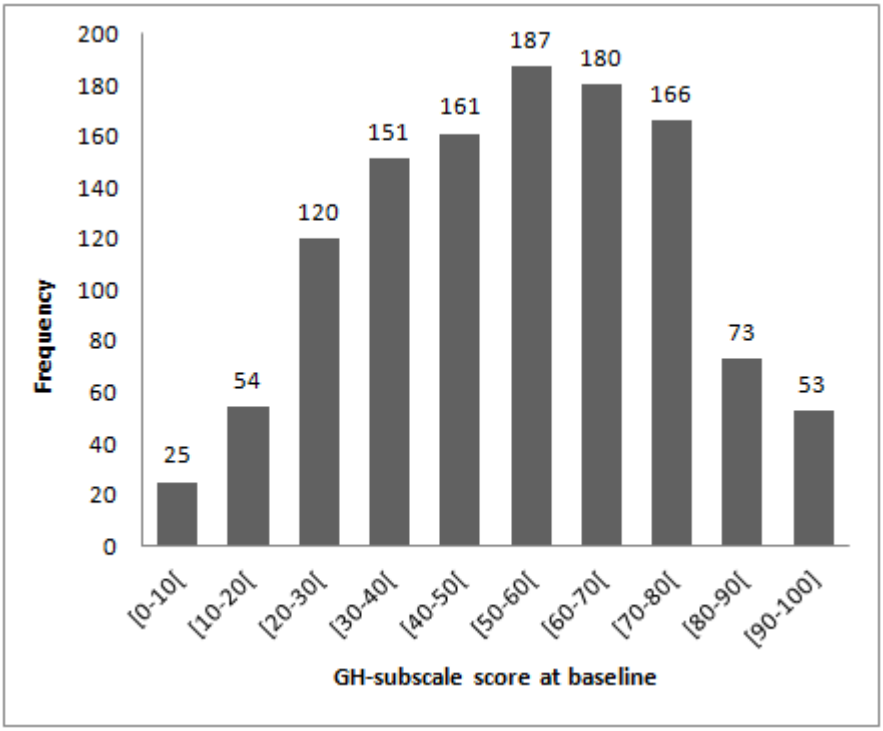

Figure 2: Box plots of the general health subscale score change from time 1 to time 2, depending on the score at time 1, in the subgroups of patients who answered “Somewhat better” (Improvement) or “Somewhat worse” (Deterioration) to the global rating of change (μ: Mean, SE: Standard Error, N: Number of patients)

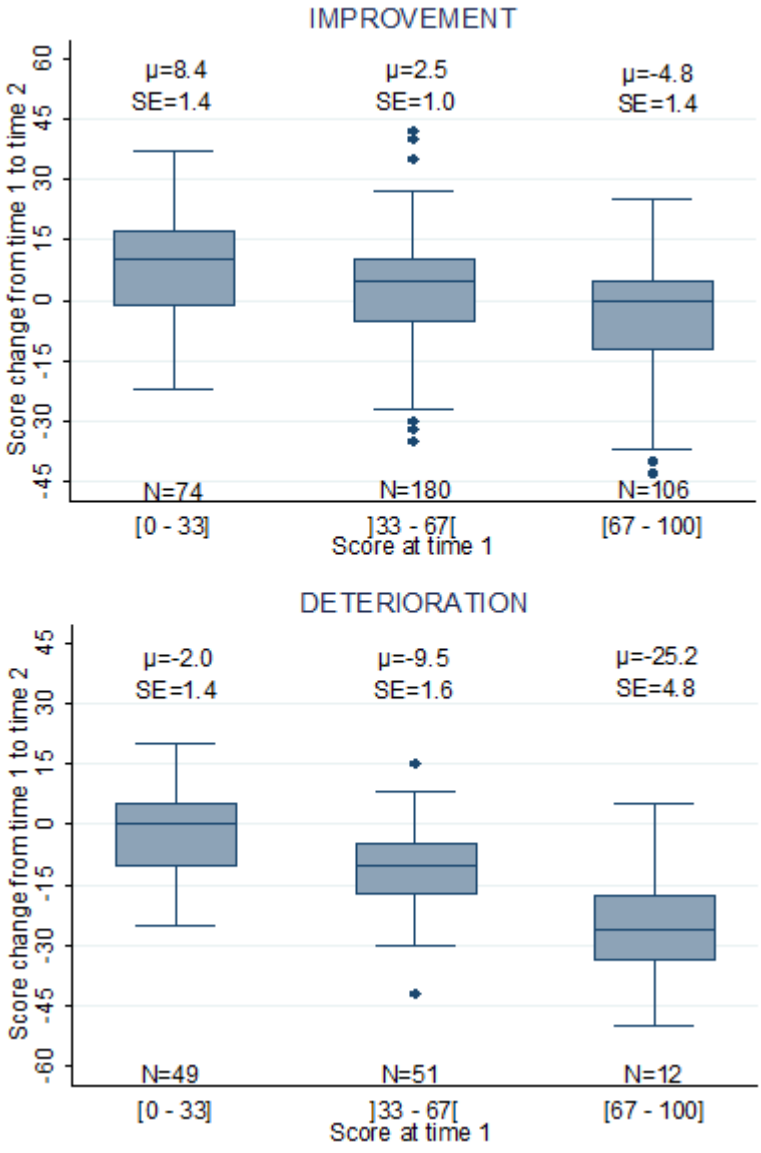

Figure 3: Expected general health subscale score depending on the trait level

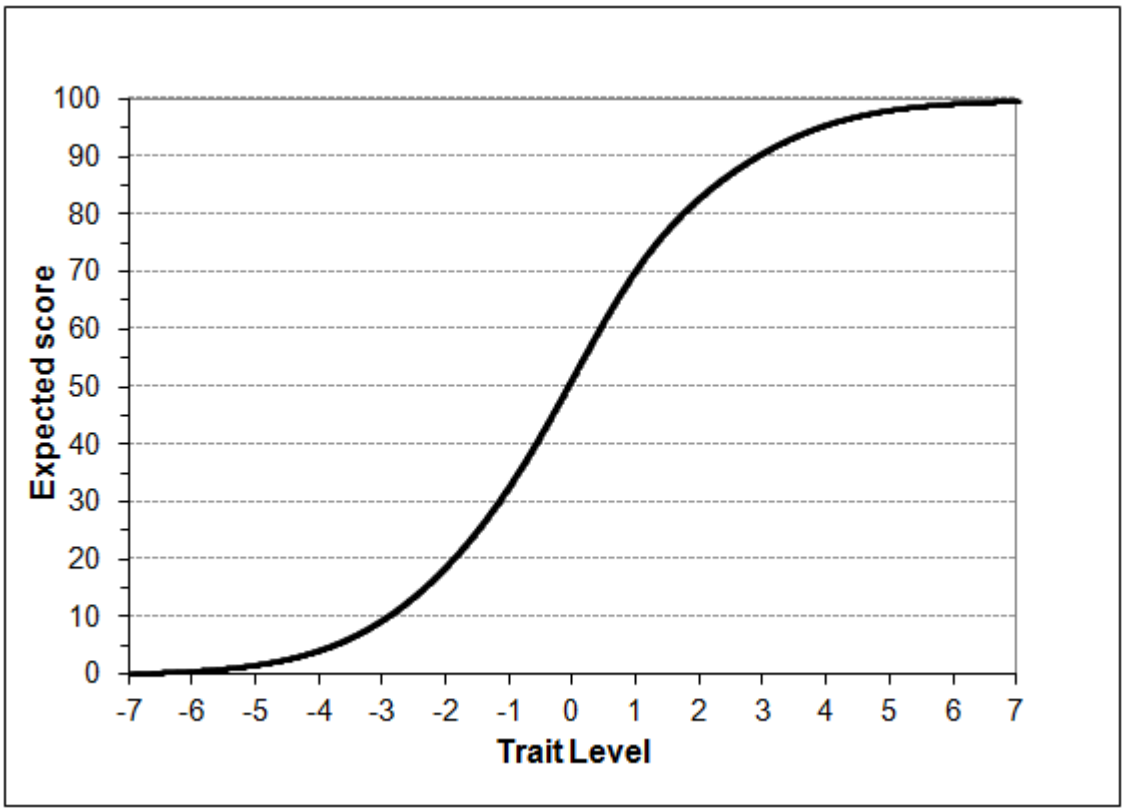

Figure 4: Minimal Clinically Important Difference determined on the Trait Level for improvement (MCID-TL= 0.0839) and deterioration (MCID-TL= -0.4806), translated on the score scale, depending on the baseline score

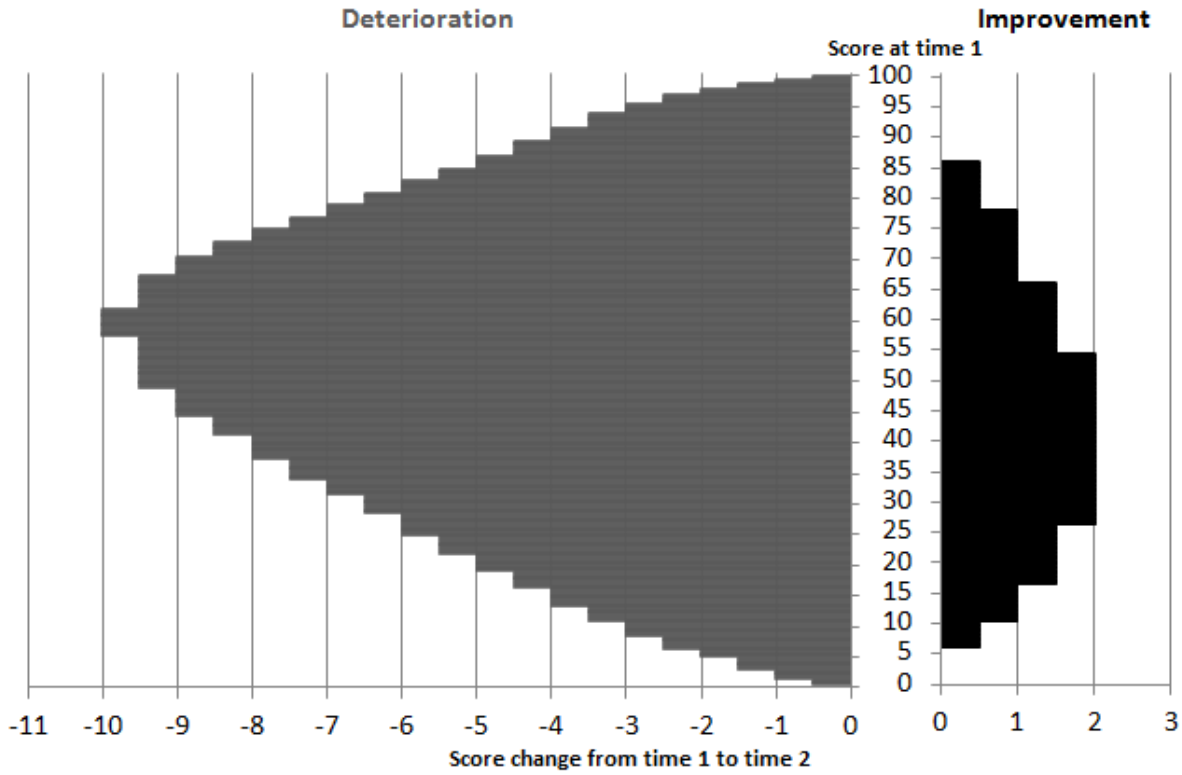

Table 1: Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) for the Minimal Clinically Important Difference determined on the score scale (MCID-Sc), on the trait level (MCID-TL) or defined as a range of values on the score scale according to the Baseline Score ($MCID\text{-}Sc_{BS}$) of the general health subscale for people who rated their health as better (Improvement) or worse (Deterioration) compared to six month ago.

		Se $[CI_{95\%}]$	Sp $[CI_{95\%}]$	PPV $[CI_{95\%}]$	NPV $[CI_{95\%}]$
Improvement	MCID-Sc	54.6% [50.7 – 58.5]	65.6% [60.9 – 70.2]	71.3% [67.2 – 75.3]	48.1% [43.9 – 52.3]
	$MCID\text{-}Sc_{BS}$	56.6% [52.7 – 60.4]	68.6% [64.0 – 73.1]	73.8% [69.8 – 77.7]	50.3% [46.1 – 54.5]
	MCID-TL	54.6% [50.7 – 58.5]	65.8% [61.2 – 70.5]	71.4% [67.4 – 75.5]	48.2% [44.0 – 52.4]
Deterioration	MCID-Sc	44.1% [35.9 – 52.2]	65.8% [61.2 – 70.5]	31.5% [25.1 – 37.9]	76.7% [72.3 – 81.2]
	$MCID\text{-}Sc_{BS}$	53.2% [45.0 – 61.3]	75.8% [71.6 – 80.0]	43.9% [36.5 – 51.3]	81.9% [78.0 – 85.9]
	MCID-TL	51.1% [42.9 – 59.2]	63.8% [59.1 – 68.5]	33.5% [27.2 – 39.8]	78.5% [74.1 – 83.0]

$CI_{95\%}$: Confidence Interval 95%

Table 2: Means of the Trait Level (TL) at each time t and t' and in each group G when a dichotomous group variable is introduced in the longitudinal mixed Partial Credit Model (for example, G indicates the response to the Global Rating of Change: 0 indicating "About the same" and 1 indicating "Somewhat better")

	Group	
	$G = 0$	$G = 1$
Time t	$\mu^{(t)}$	$\mu^{(t)} + \alpha + \beta^{(t)}$
Time t'	$\mu^{(t\prime)}$	$\mu^{(t\prime)} + \alpha + \beta^{(t\prime)}$
Mean change of the TL (Time effect)	$\mu^{(t\prime)} - \mu^{(t)}$	$\left(\mu^{(t')} - \mu^{(t)}\right) + (\beta^{(t')} - \beta^{(t)})$

Supplementary material

4.1 The Partial Credit Model

The PCM is an IRT model for polytomous data belonging to the Rasch models, in which the probability of a response y_l to an item j ($j = 1, \dots, J$) with l categories ($l = 1, \dots, m_j$) for the subject i ($i = 1, \dots N$) is a function of the subject's TL (denoted θ_i). It can be written:

$$P(Y_{ij} = y_l/\theta_i, \delta_{jl}) = \frac{exp\left(y_l\theta_i - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i - \sum_{l=1}^{c}\delta_{jl}\right)}$$

where δ_{jl} is the item parameter associated to the response category l of the item j [61]. The relationship between the TL and the expected score on the scale, $E(S)$, can be calculated using the following equation [41]:

$$E(S) = \sum_{j=1}^{J}\sum_{l=1}^{m_j} y_l \times P(Y_j = y_l/\theta, \delta_{jl})$$

4.2 The Longitudinal Mixed Partial Credit Model

If θ is considered as a random variable having, for example, a normal distribution $N(\mu, \sigma^2)$, the PCM is a mixed-effects logistic model in which the parameters to be estimated are μ (the TL mean in the sample), σ (the TL standard error in the sample) and δ_{jl} (the item parameters)[62]. For repeated measurements, a longitudinal form of the mixed-effects PCM has been developed as it has yet been done for the Rasch model by Blanchin et al [53, 63]. The probability of a response in the category y of an item j at time t ($t = 1, \dots T$) can be written as:

$$P(Y_{ij}^{(t)} = y^{(t)}/\theta_i^{(t)}, \delta_{jl}) = \frac{exp\left(y^{(t)}\theta_i^{(t)} - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} - \sum_{l=1}^{c}\delta_{jl}\right)}$$

The item parameters δ_{jl} are assumed to be constant within time assessments (i.e. measurement invariance is assumed) and can be estimated or are considered as known and fixed in the model. The other parameters to be estimated are the $\mu^{(t)}$ parameters (the TL mean in the sample at time t), $\sigma^{(t)}$ (the TL standard error in the sample at time t) and $\sigma^{(tt')}$ (the covariance between $\theta^{(t)}$ and $\theta^{(t')}$) with $t \neq t'$. The distribution of the TL is assumed to be a multinormal distribution of dimension T. The mean change over time of the construct measured by the questionnaire in the entire sample can be evaluated by the evolution of $\mu^{(t)}$ across each time t.

Dichotomous group variables (G with realisation g_i for the subject i) can be introduced in this model to be able to assess this time effect in various groups in the sample:

$$P(Y_{ij}^{(t)} = y^{(t)} | \theta_i^{(t)}, \delta_{jl}, \alpha, \beta^{(t)}) = \frac{exp\left(y^{(t)}\theta_i^{(t)} + \boldsymbol{\alpha g_i} + \boldsymbol{\beta^{(t)} g_i} - \sum_{l=1}^{y}\delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} + \boldsymbol{\alpha g_i} + \boldsymbol{\beta^{(t)} g_i} - \sum_{l=1}^{c}\delta_{jl}\right)}$$

With the identifiability constraint $\beta^{(1)} = 0$, in this model, α represents the mean difference between the two groups at first time ($t = 1$) and $\alpha + \beta^{(t)}$ represents the mean difference between the two groups at time t. Means of the TL at each time t and t', and in each group are indicated in **table 2**. If the variable G indicates the response to the GRC (0 indicating "About the same" and 1 indicating "Somewhat better"), the MCID-TL for improvement is equal to the mean change of the TL from time t to time t' in the group $G = 1$, that is $(\mu^{(t')} - \mu^{(t)}) + (\beta^{(t')} - \beta^{(t)})$.

[Table 2 near here]