Establishing Thresholds for Meaningful Within-individual Change Using Longitudinal Item Response Theory

Jakob Bue Bjorner (jbjorner@qualitymetric.com)
QualityMetric https://orcid.org/0000-0001-7033-8224
Berend Terluin
University of Amsterdam: Universiteit van Amsterdam
Andrew Trigg
Mapi Values UK: Adelphi Values
Jinxiang Hu
University of Kansas Medical Center
Keri J.S. Brady
Sanofi-Aventis US LLC
Pip Griffiths
IQVIA Solutions UK Ltd

Research Article

Keywords: threshold, meaningful within-individual change (MWIC), patient-reported outcome measures (PROM), Graded Response LIRT model

Posted Date: July 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-371137/v1

ESTABLISHING THRESHOLDS FOR MEANINGFUL WITHIN-INDIVIDUAL CHANGE USING LONGITUDINAL ITEM RESPONSE THEORY

Jakob Bue Bjorner[1,2], jbjorner@qualitymetric.com, [1]QualityMetric Incorporated, LLC, [2] University of Copenhagen

Berend Terluin[3], b.terluin@amsterdamumc.nl, [3] Amsterdam University Medical Centers

Andrew Trigg[4], andrew.trigg@adelphivalues.com , [4] Adelphi Values, Bollington

Jinxiang Hu[5], jhu2@kumc.edu , [5] University of Kansas Medical Center

Keri J. S. Brady[6], keri.brady@sanofi.com, [6] Sanofi

Pip Griffiths[7], pip.griffiths@iqvia.com, [7] IQVIA

Abstract: 245 words (250 Max)

Text: 3940 words (4000 Max)

Declarations

Funding Funded by the institutions of the participating authors

Conflicts of interest/Competing interests No conflicts of interest

Availability of data and material (data transparency) Data available through Journal Website

Code availability (software application or custom code) Code available through Journal Website

Authors' contributions: JBB suggested the method, conducted data analyses, participated in writing the first draft, performed revisions, and approved the final version. BT provided additional suggestions for the method, conducted data analyses, participated in writing the first draft, commented on revisions, and approved final version. AT provided additional suggestions for the method, commented on revisions, and approved final version. JH conducted data analyses, commented on revisions, and approved final version. KB provided additional ideas, commented on revisions and approved final version. PG provided additional suggestions for the method, commented on revisions, and approved final version.

ESTABLISHING THRESHOLDS FOR MEANINGFUL WITHIN-INDIVIDUAL CHANGE USING LONGITUDINAL ITEM RESPONSE THEORY

Abstract

PURPOSE: Thresholds for meaningful within-individual change (MWIC) are useful for interpreting patient-reported outcome measures (PROM). Transition ratings (TR) have been recommended as anchors to establish MWIC. Traditional statistical methods for analyzing MWIC such as mean change analysis, receiver operating characteristic (ROC) analysis, and predictive modeling ignore problems of floor/ceiling effects and measurement error in the PROM scores and the TR item. We present a novel approach to MWIC estimation for multi-item scales using longitudinal item response theory (LIRT).

METHODS: A Graded Response LIRT model for baseline and follow-up PROM data was expanded to include a TR item measuring latent change. The LIRT threshold parameter for the TR established the MWIC threshold on the latent metric, from which the observed PROM score MWIC threshold was estimated. We compared the LIRT approach and traditional methods using an example data set with baseline and three follow-up assessments differing by magnitude of score improvement, variance of score improvement, and baseline-follow-up score correlation.

RESULTS: The LIRT model provided good fit to the data. LIRT estimates of observed PROM MWIC varied between 3 and 4 points score improvement. In contrast, results from traditional methods varied from 2 points to 10 points - strongly associated with proportion of self-rated improvement. Best agreement between methods was seen when approximately 50% rated their health as improved.

CONCLUSION : Results from traditional analyses of anchor-based MWIC are impacted by study conditions. LIRT constitutes a promising and more robust analytic approach to identifying thresholds for MWIC.

Introduction

Patient-Reported Outcome Measures (PROMs) bring the necessary patient perspective into the evaluation of treatment outcomes – assessing constructs such as pain, physical functioning, or depression. The PROM score (often the sum of responses to several items) provides an estimate of the construct at individual or group levels. When PROMs are administered longitudinally (e.g., pre- and post-treatment), score differences estimate individual or mean group change. A problem in interpreting PROM scores is that they lack intrinsic meaning. It is difficult to tell if a given PROM change score (of e.g., 10 points on a 100 points scale) represents a substantial or meaningful change. To assist PROM users in the interpretation of change scores, the concept of the "minimal important change" (MIC, also denoted "minimal important difference", "minimal clinically important difference" or "meaningful change threshold") was introduced [1, 2]. Subsequent discussions have distinguished changes between groups, within groups, or within individuals [3, 4]. This paper considers thresholds for meaningful within-individual change (MWIC).

Methods to establish MIC and MWIC thresholds fall in two major categories: distribution-based and anchor-based methods [5, 6]. Anchor-based methods employ an external criterion (an anchor), often a patient-reported global rating of change, which we denote a transition rating (TR) [7]. We will focus on anchor-based methods, because they incorporate the patient's perspective of meaningful change [8]. Both PROM improvement and deterioration may be of interest, but for the sake of simplicity, we will focus on MWIC thresholds for improvement. The methods are equally applicable to thresholds for deterioration.

Traditional statistical methods to estimate thresholds for MIC or MWIC based on TRs include the mean change method [9], the receiver operating characteristic (ROC) method [10], and the predictive modeling method [11]. These methods have a number of drawbacks. The mean change method simply takes the mean PROM change score in the group rating their perceived change as small but meaningful (e.g. "a little better") as the MWIC. It is unclear whether this group average can serve as a threshold for MWIC, and it is typically higher than thresholds derived by other analyses [12]. Also, measurement error in the TR item may bias the selection of respondents with small but meaningful improvement. The ROC method seeks to identify the PROM change score threshold that optimally classifies patients into improved (e.g. indicating "a little better" or "much better" on the TR item) from not-improved (e.g. indicating "unchanged" or worse) patients. Often the threshold that minimizes the absolute difference between sensitivity and specificity is chosen as MWIC threshold, ensuring the least misclassification [13]. However, the ROC-based MWIC has been shown to be imprecise, due to measurement error in the PROM change score [11], and it is biased by the proportion of improved patients [14]. The predictive modeling method uses logistic regression

modeling to identify the PROM change score that is equally likely to occur in the improved and not-improved groups (i.e., the PROM change score with a likelihood ratio of 1) [11, 15]. This approach assumes a linear effect of the PROM change score on the log odds of TR response, which may not be realistic in case of floor or ceiling effects. The approach is also biased by the proportion of improved patients, although procedures have been proposed to adjust for this bias [14].

In this paper we will introduce a new approach to estimate an MWIC, based on longitudinal item response theory (LIRT) [16]. Item response theory (IRT) refers to a group of mathematical models describing the relationship between the responses to the items of a questionnaire and the constructs the questionnaire is purported to measure [17, 18]. The constructs are modeled as latent (i.e., not directly observable) variables. The LIRT model is a multidimensional IRT model [19] that allows item responses to be influenced by latent variables at different time points. We will use a LIRT model to estimate an MWIC threshold based on responses to a TR item. First, we show how the MWIC threshold can be defined and estimated on the latent scale. Second, we use the LIRT model to estimate the MWIC for the observed PROM change score. Third, we discuss how to evaluate LIRT model assumptions. Fourth, we evaluate the new approach using the dataset provided for this special issue and we compare our LIRT-based MWIC results with the results of the traditional approaches.

Methods

Defining and Estimating MWIC Thresholds on the Latent Scale

We start by developing a statistical model for a situation where a set of items measuring a particular health construct has been administered at baseline (time 1) and follow-up (time 2). Also, a TR item assessing change in the same construct has been administered at time 2. For simplicity of discussion, we assume that responses to the TR item have been collapsed into 1: meaningful improvement ("minimally improved" to "much improved") and 0: no meaningful improvement ("much worse" to "no change"). The model can easily be extended to cover all the response categories on the TR item. The IRT model for the TR item can be specified in the following way (see Appendix for details):

$$ln\left(\frac{P(TR_j=1)}{P(TR_j=0)}\right) = \alpha_{TR}(d\theta_j - \beta_{TR}) \quad (1)$$

where $ln\left(\frac{P(TR_j=1)}{P(TR_j=0)}\right)$ is the log-odds of person *j* scoring '1' on the TR, $d\theta_j$ is the latent change from time 1 to time 2 for person *j*, β_{TR} is the MWIC threshold and α_{TR} is a so-called discrimination parameter reflecting the TR item's measurement precision (see Appendix). This IRT model is illustrated in Figure 1, which shows the probability of indicating a meaningful improvement on the TR item for different levels of latent health change. The MWIC threshold is the level of health improvement where the probability of answering 1 (meaningful improvement) equals the probability of 0 (no meaningful improvement). The model in Figure 1 has an MWIC threshold of 0.56.

[Figure 1 about here]

The IRT model in equation 1 and Figure 1 is a standard two-parameter IRT model except that the latent variable is health change and not health status. However, to estimate the model, we need more indicators of health change. To achieve these indicators, we can embed the IRT model for the TR item within an LIRT model.

Such an LIRT model is illustrated in Figure 2. The TR item loads on the latent health states at both time 1 (θ_1) and time 2 (θ_2). We impose the restriction that the discrimination parameter for θ_1 is of the same magnitude but opposite sign as the discrimination parameter for θ_2 (please see Appendix). With this restriction, the item parameters for the TR item (α_{TR}, β_{TR}, equation 1) can be estimated from the LIRT model. The bottom half of Figure 2 (in grey) illustrates a possible model extension: it is plausible that responses to the same item at two different time points are locally dependent. This local dependence can be modeled by including a latent variable for each item capturing the local dependence across time - sometimes also called a two-tiered model [20]. These LIRT models can be estimated with modern software for multidimensional IRT (e.g. Mplus [21] or the mirt package in R [22]) and β_{TR} can be estimated directly (in Mplus) or indirectly (in mirt).

[Figure 2 about here]

Estimating MWIC thresholds on the Observed PROM Scale

For scales that are scored by IRT estimation of the latent score θ (e.g. scales from the PROMIS project), the MWIC threshold is provided by the β_{TR} estimate. However, most PROMs are scored by calculating the sum of the items (which may be linearly transformed to a metric ranging from 0 to 100 or to a T-score metric). To estimate the MWIC for the observed PROM change score, we use a 4-step Monte-Carlo estimation procedure based on the parameters from the LIRT model: 1. Simulate the distribution of θ_1 using the mean and variance from the above LIRT model (typically fixed at 0

and 1). 2. Calculate θ_2 scores based on the assumption that all respondents had a latent improvement ($d\theta_j$) of exactly β_{TR} (the MWIC threshold on the latent scale). 3. Simulate item responses at T1 and T2 based on θ_1, θ_2, and the estimated LIRT item parameters; calculate the "observed" PROM scores and PROM change scores. 4. Use the median PROM change score as the MWIC threshold for the observed PROM change score. We use the median rather than the mean since the distribution of PROM change scores may be skewed if the mean baseline score is close to the floor or the ceiling of the scale. However, we also present the mean PROM change score for comparison.

Model Assumptions and Test of Model Fit

The described LIRT model makes five major assumptions that can all be tested: 1. At each time point, the IRT model should provide good fit to PROM items. This assumption can be tested by standard IRT methods, e.g. by a generalization of the S-X^2 test [23] to polytomous items. Misfit may lead to use of a more general IRT model (such as the nominal categories model [24]) for some items. 2. The PROM item parameters are assumed to be the same for time 1 and time 2 (i.e. no item response shift). This assumption can be tested for one item at the time using a likelihood ratio test comparing a model constraining the item parameters for that item over time with a model without these constraints. In case of significant difference for some items, the item parameters for these items can be allowed to differ over time. 3. The simplest form of the LIRT model assumes local item independence across time. This can be tested by comparing models with and without local dependence using a likelihood ratio test. The magnitude of local dependence can be evaluated by the discrimination parameters for the local dependence latent variables. Significant local dependence can be included in the model. 4. The discrimination parameter for the TR item on θ_1 is assumed to be of the same magnitude but opposite sign as the discrimination parameter for the TR item on θ_2. This can be tested by comparing models with and without constraints on the discrimination parameter for the TR item. Differences in the magnitude of the discrimination parameter for θ_1 and θ_2 can be caused by present state bias (see Terluin et al, under review). 5. Finally we assume that the LIRT model provides good fit for the TR item. This may be evaluated by estimating the expected proportion of positive (1) responses to the TR item for different levels of the observed PROM change score. These expected proportions can be derived by simulation based on the estimated model parameters. These expected proportions can then be compared with the observed proportions of positive answers (similar to the approach of Orlando and Thissen [23]).

Analysis of Example Dataset

The example data set was created for this special issue by simulation. The details of the simulations were unknown to the analyst for the current study. The dataset contains 2000 subjects' responses to 12 hypothetical items from the Simulated Disease Questionnaire (SDQ-12). Each item has 4 response categories. Items were originally scored such that high scores indicate poor health; we reverse-coded the original scores to be consistent with our discussions, so high scores indicate good health. Scores are available for four time points: baseline and follow-up 1, follow-up 2, and follow-up 3. Additionally, responses to a seven category TR item are available at follow-up 1, 2, and 3. For the current analyses, we have collapsed responses to 1: Much improved, Moderately improved, Minimally improved, and 0: No change, Minimally worse, Moderately worse, Much worse. The MWIC threshold was simulated to be the same across follow-up time points.

Our analyses established an MWIC threshold for changes from baseline to follow-up 1, 2, and 3, respectively. For each follow-up, we estimated MWIC on the latent metric (see Appendix for Mplus and mirt code) and evaluated model assumptions according to the procedures described above. Then we derived MWIC estimates for the observed PROM scale and compared our results with traditional analyses: mean change, ROC, and adjusted predictive modeling [14]. To obtain 95% confidence intervals (CIs) for the LIRT MWIC estimates on the observed metric, we derived observed PROM MWIC estimates based on the lower and upper bound β_{TR} values, holding all other parameters in the model constant. For mean change analysis, 95% CIs were estimated directly, while for ROC analyses and adjusted predictive modeling, 95% CIs were achieved by empirical bootstrap (1000 samples). For ROC analyses, we choose as MWIC threshold the value that minimized the absolute difference between sensitivity and specificity [13].

Results

The top part of Table 1 shows descriptive information for each of the three follow-up time points. Based on the PROM change score, the sample had no significant mean change at follow-up 1, but had mean improvements of 0.19 SD at follow-up 2, and 0.58 SD at follow-up 3. Similarly, the percentage of respondents, who themselves reported that they had improved, increased over time from 34.5% at follow-up 1, to 51.1% at follow-up 2, and 72.0% at follow-up 3. The polychoric correlations between the TR item and the PROM change score were between 0.57 and 0.61; far above the threshold 0.30, which is sometimes used to assess whether a TR item is sufficiently associated with PROM score change to make MWIC estimation meaningful [8].

The next section of Table 1 presents parameter estimates from LIRT modeling of each of the 3 follow-up times compared to baseline. Estimates of the mean θ_2 values (equal to mean $d\theta$ since mean of $\theta_1$is zero) supported the results of analyses of PROM change scores that no significant mean improvement was seen at follow-up 1, but increasing improvements were seen at follow-up 2 and 3. The θ_2 SD estimates ranged from 1.2 at follow-up 1 to 1.7 at follow-up 3. Even larger increases were seen in the $d\theta$ SD estimates (which were calculated from θ_1 SD, θ_2 SD, and the θ_1*θ_2 correlation). The baseline*follow-up correlation decreased from 0.58 (at follow-up 1) to 0.18 (at follow-up 3). The TR item discrimination parameter varied between 0.94 (follow-up 3) and 1.63 (follow-up 1). Finally, the TR item threshold parameter estimates varied between 0.46 (follow-up 3) and 0.56 (follow-up 1). Figure 1 in the introduction illustrates the LIRT model for the TR item at follow-up 1 (i.e. α_{TR} = 1.63 and β_{TR}= 0.56).

In tests of model fit at each time point, the generalized S-X^2 item level fit tests did not show significant misfit. Out of 48 item-level fit test (12 items at 4 time points), the lowest P-value was 0.027 and only two P-values were below 0.05. We did not find any indication of significant item response shift. Modeling of local item dependence across time did not improve model fit at any follow-up time point and was therefore not included in the final models. Equality of the discrimination parameters for the TR item on $\theta_1$and θ_2 was supported at all three follow-up times. The largest difference was found at follow-up time 3 where separate estimation yielded α_{TR1} = -1.05, α_{TR2}= 0.92 (likelihood ratio test for significant difference = 3.37, P = 0.066).

Figure 3 shows plots of expected and observed proportions of self-rated improvement along with the S-X^2 test of fit, which was used to assess TR item fit. The model fit is acceptable at each follow-up time. Although the P-value is 0.04 at follow-up 3, there are no indications of systematic departures from the predicted model.

[Figure 3 about here [25]]

Figure 4 shows that in the θ range from -1 to 1.5, the expected PROM score function is almost linear. At baseline (where mean of θ_1 = 0 and SD θ_1 = 1) very few respondents are close to the floor or ceiling of the scale (indicated by a flat expected PROM score function).

[Figure 4 about here]

Estimates of MWIC based on LIRT varied between 3 and 4 points score improvement (Table 1). Estimates of MWIC based on mean change, ROC analyses, and predictive modeling were lower than the LIRT based estimate for follow-up 1 (range 1.78-2.74) but higher than the LIRT based estimate

for follow-up 3 (range 6.71 to 10). At follow-up 2, all four methods agreed on an MWIC estimate close to 3.

Discussion

We introduced a new method to estimate the MWIC of a PROM, based on longitudinal IRT. This method was compared to traditional methods using a simulated data set of PROM and TR item responses. Whereas the new method identified MWIC values between 3 and 4 across the three follow-up points, the traditional MWIC estimates showed much more variability, and bias related to the proportion of improved patients. That the ROC-based MWIC is biased by the proportion improved is well known [14, 26]. The adjusted MWIC is supposed to adjust for this bias [14], but the present findings suggest that the adjusted MWIC is also biased by the proportion improved. The mean change MWIC appeared also to be biased by the proportion improved, a finding that has not been described before. We hypothesize that the most likely reason for the bias of traditional methods is that they do not adequately handle measurement error in the TR item and in the PROM change score. Measurement error in the TR item implies that a proportion of persons who rate themselves as improved would not provide the same rating if asked again. If true improvement (defined as $d\theta \geq \beta_{TR}$) only occurs in a small part of the sample (notably below 50%), a large proportion of those who rate themselves as improved provide this rating due to measurement error, thus the TR item overestimates the true proportion of improved respondents. The persons providing a TR rating of "Minimally improved" due to measurement error is likely to have a low PROM change score, thus biasing the MWIC threshold estimate of traditional methods downwards. In contrast, if the proportion of true improvement is large (notably above 50%), measurement error will cause TR to underestimate the true proportion of improved persons, thus biasing the MWIC threshold estimate upwards. This is a problem for techniques such as mean change, ROC analyses, and predictive modeling, which assume that the TR item is a gold standard without error.

Advantages of the new method

One of the advantages of our suggested approach is that the expanded LIRT model can handle both measurement error in the TR item and in the PROM change score. For follow-up 1, the α_{TR} parameter of 1.63 implies an item reliability of 0.49 (see Appendix). However, even if the reliability of the TR item is less than perfect, the LIRT model provides unbiased estimates of β_{TR} (and thus the MWIC threshold for the PROM change score). Similarly, while the PROM change score will

have measurement error, the independent variable in our model, $d\theta$, is not subject to error if the model is correctly specified.

Limitations and future research

While the LIRT model provides a probability of rating 'improved' for each individual, the probability will never be 0 or 1, even if the true change ($d\theta$) could be known. Two factors may contribute to this lack of certainty: 1. Persons may differ in their individual threshold for MWIC [14]. Thus, the MWIC threshold estimated by the model can be seen as the mean of individual thresholds (see Appendix). 2. There may be individual differences in the way the TR item is interpreted and answered (what we would typically call measurement error). These two factors cause the reliability of the TR item to be less than 1 – often below 0.5 as in the current study. A further source of random variation is the measurement error of the PROM change score as a measure of the true change ($d\theta$). Thus, while the prediction is valid for each person, it is also uncertain. We believe that this uncertainty in the model prediction represents real uncertainty in a classification based on a single item – as illustrated in Figure 3. Given this uncertainty, strategies could be used to identify persons with a higher probability of being meaningfully improved. One strategy would be to require that patients – in addition to being more likely than not to rate their improvement as meaningful – should have a high probability of truly being improved. In classical psychometrics this idea is used in the reliable change index (RCI) [27]. For our example, the RCI with 80% confidence interval is +/- 6.33. Another strategy would be to require a threshold with higher probability of rating the change as meaningful. For example, if requiring at least 75% probability of rating the change as meaningful, equation 1 and results from follow-up 1 leads to a $d\theta$ threshold value of 1.23 and an MWIC of 8 on the PROM change scale. However, any of these attempts to identifying a group with highly likely meaningful change will cause many persons with meaningful improvement to be misclassified as not improved. An advantage of using the 50% probability criterion is that, on the group level, if X% of the group has improved by more than the MWIC, it is safe to say that about X% has experienced a meaningful change.

The LIRT-model requires a number of assumptions to be met. As we have shown, these assumptions can be evaluated and the LIRT model can often be revised to accommodate misfit. An important assumption is that the transition rating is a valid indicator of the latent change. This can be tested by evaluating the item fit and by looking at the magnitude of the transition rating discrimination parameter. However, some TR validity problems may not be easily detected: e.g. social desirability

bias or impact of other questions (order effects). These potential sources of error are general to all analytic methods and emphasize the wisdom of using multiple anchors and methods to establish a threshold for MWIC.

We have assumed that the discrimination parameter for the TR item on θ_1 is of the same magnitude but opposite sign as the discrimination parameter for the TR item on θ_2. However, as discussed above, transition ratings might be more affected by the present (i.e., follow-up) state than the change. Then we might find that the discrimination parameter for the TR item on θ_2 would be numerically larger than the discrimination parameter for the TR item on θ_1. More work is needed to understand the best approach to MWIC estimation in this situation.

The translation from a theta change MWIC to a PROM scale change MWIC is not straightforward as the relationship between the theta and the PROM scale score is non-linear (Figure 4). In this paper we used a Monto-Carlo approach and chose the median PROM scale change for subjects who improved by exactly the MWIC on the latent scale. An alternative choice would be taking the mean of the PROM change score instead of the median. The approaches agree for the current analyses (see Table 1), since very few observations are at the floor or the ceiling and the change score distribution is fairly symmetrical. However, if baseline samples have highly skewed PROM scores and floor or ceiling effects, the approach to transforming MWIC threshold from the latent scale to the PROM change scale may matter more. Further work is needed to establish the best way to translate the theta change MWIC into a PROM scale change MWIC in case of floor or ceiling effects.

Conclusions

We have presented a new approach to estimating MWIC based on TR items. This approach, using LIRT, gave consistent results in the data sets provided for this special issue. In contrast, results from traditional analyses of anchor-based MWIC were impacted by study conditions. Thus, LIRT constitutes a promising and more robust analytic approach to identifying thresholds for MWIC in multi-item scales.

References

1. Guyatt, G. H., Walter, S., & Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *J Chron Dis,* 40, 171–178.
2. Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of Health Status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials,* 10, 407–415.
3. Coon, C. D., & Cappelleri, J. C. (2016). Interpreting change in scores on patient-reported outcome instruments. *Therapeutic innovation & regulatory science,* 50(1), 22–29.
4. Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research,* 27(1), 33–40.
5. King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res,* 11, 171–184.
6. Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol.,* 56(5), 395–407.
7. Guyatt, G. H., Norman, G. R., Juniper, E. F., & Griffith, L. E. (2002). A critical look at transition ratings. *Journal of Clinical Epidemiology,* 55(9), 900–908.
8. Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., et al. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ (Clinical research ed.)*. doi:10.1136/bmj.m1714.
9. Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K.-K. (2005). Evaluating the statistical significance of health-related quality-of-life change in individual patients. *Evaluation & the Health Professions,* 28(2), 160–171.
10. Deyo, R. A., & Centor, R. M. (1986). Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of Chronic Diseases,* 39(11), 897–906.
11. Terluin, B., Eekhout, I., Terwee, C. B., & de Vet, Henrica CW (2015). Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. *Journal of Clinical Epidemiology,* 68(12), 1388–1396.
12. Bjorner, J. (2019). PNS320 ANALYSIS OF MINIMAL IMPORTANT CHANGE THROUGH ITEM RESPONSE THEORY METHODS. *Value in Health,* 22, S818.
13. Farrar, J. T., Young Jr, James P, LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain,* 94(2), 149–158.
14. Terluin, B., Eekhout, I., & Terwee, C. B. (2017). The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. *Journal of Clinical Epidemiology,* 83, 90–100.
15. Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *bmj,* 329(7458), 168–169.
16. Wang, C., & Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics,* 45(3), 339–368.
17. Van der Linden, Wim J (2018). *Handbook of item response theory, three volume set* : CRC Press.
18. Embretson, S., & Reise, S. P. (2009). *Item Response Theory for Psychologists* (2nd edn). Mahwah, NJ: Lawrence Earlbaum Associates, Publishers.
19. Bonifay, W. (2019). *Multidimensional item response theory* : Sage Publications.

20. Cai, L. (2010). A Two-Tier Full-Information Item Factor Analysis Model with Applications. *Psychometrika,* 75(4), 581–612.
21. Muthen, B. O., & Muthen, L. (2017). *Mplus User's Guide* (8th edn). Los Angeles: Muthen & Muthen.
22. Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., et al. (2015). mirt: Multidimensional item response theory. *Computer Software) http://CRAN. Rproject. org/package= mirt*.
23. Orlando, M., & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement,* 24(1), 50–64.
24. Bock, R. D. (1997). The Nominal Categories Model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 3–50). Berlin: Springer.
25. Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika,* 26(4), 404–413.
26. Terluin, B., Griffiths, P., van der Wouden, Johannes C, Ingelsrud, L. H., & Terwee, C. B. (2020). Unlike ROC analysis, a new IRT method identified clinical thresholds unbiased by disease prevalence. *Journal of Clinical Epidemiology,* 124, 118–125.
27. Jacobson, N. S., & Truax, P. (1991). Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *J Consult Clin Psychol,* 59(1), 12–19.
28. Takane, Y., & Leeuw, J. de (1987). On The Relationship Between Item Response Theory And Factor Analysis Of Discretized Variables. *Psychometrika,* 52(3), 393–408.
29. Muthen, B. O., & Christoffersson, A. (1981). Simultaneous Factor Analysis of Dichotomous Variables in Several Groups. *Psychometrika,* 46, 407–419.
30. Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics,* 19, 293–295.
31. Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Qual.Life Res.,* 16 Suppl 1, 95–108.
32. Samejima, F. (1997). Graded response model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). Berlin: Springer.

Figures and tables: ESTABLISHING THRESHOLDS FOR MEANINGFUL WITHIN-INDIVIDUAL CHANGE USING LONGITUDINAL ITEM RESPONSE THEORY

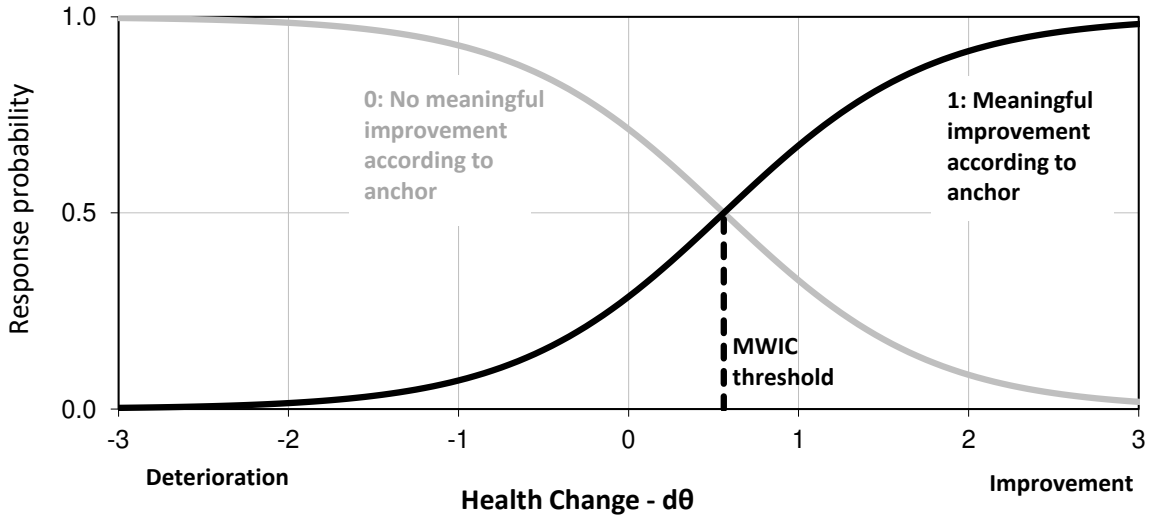

Figure 1. Probability of rating a health improvement as meaningful according to the IRT model for a transition rating item

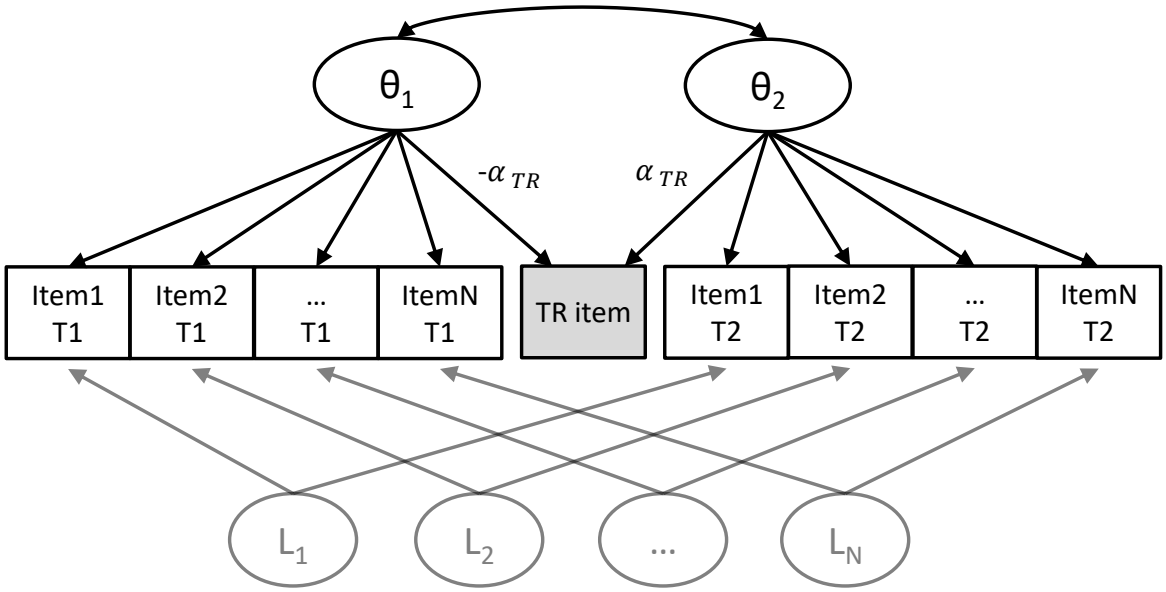

Figure 2. A longitudinal IRT model including a transition rating (TR) item.

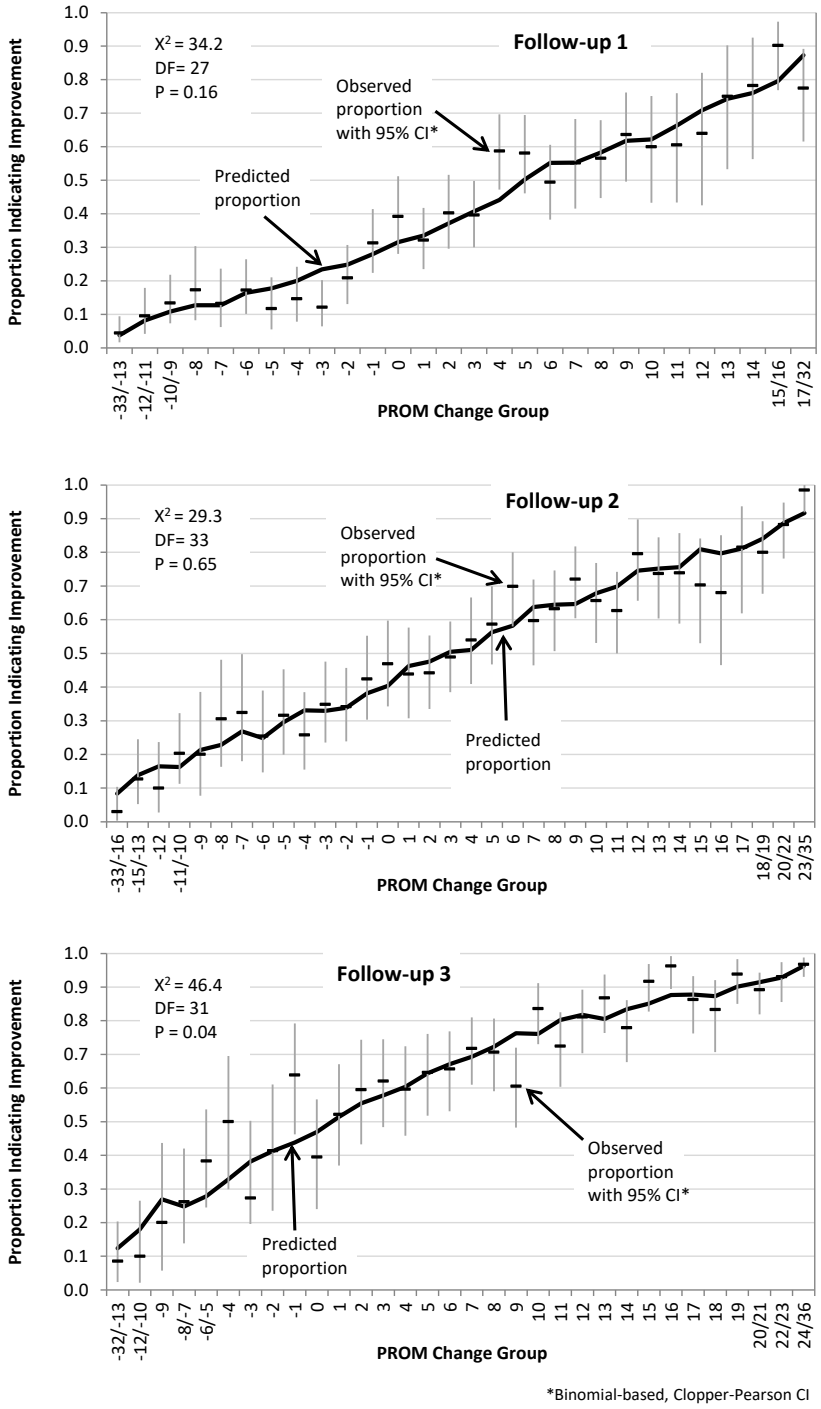

Figure 3. Testing the assumption of transition rating (TR) item fit: Plots of expected and observed proportions of respondents with different PROM change levels indicating improvement on the transition rating item. Full line: expected proportions of self-rated improvement. Dots with horizontal lines: observed proportions with 95% CI according to a binomial model [25]. PROM change scores are collapsed at the extremes so that all expected cell frequencies are > 5.

Figure 4. Plots of expected scores on the PROM across levels of θ.

Table 1. Descriptive statistics and parameter estimations from analysis of MWIC

	Follow-up 1		Follow-up 2		Follow-up 3	
	Est	(95% CI)	Est	(95% CI)	Est	(95% CI)
Descriptive information						
dPROM mean (raw score)	-0.20	(-0.56: 0.17)	3.45	(2.99: 3.90)	10.24	(9.79: 10.7)
dPROM mean (ES)	-0.01	(-0.03 : 0.01)	0.19	(0.17 : 0.22)	0.58	(0.55 : 0.60)
% improved	34.5%	(32.4% : 36.5%)	51.1%	(48.9% : 53.3%)	72.0%	(70.0% : 73.9%)
TR*dPROM polychoric correlation	0.57	(0.53 : 0.61)	0.58	(0.54 : 0.62)	0.61	(0.57 : 0.65)
LIRT parameter estimates						
θ_1 mean[1]	0		0		0	
θ_1 SD[1]	1		1		1	
θ_2 mean/$d\theta$ mean[2]	-0.02	(-0.08 : 0.03)	0.57	(0.50 : 0.65)	1.95	(1.84 : 2.07)
θ_2 SD	1.20	(1.15 : 1.26)	1.45	(1.38 : 1.53)	1.70	(1.61 : 1.80)
$d\theta$ SD	1.03	(0.98 : 1.08)	1.47	(1.40 : 1.54)	1.81	(1.72 : 1.91)
θ_1*θ_2 Correlation	0.58	(0.54 : 0.61)	0.33	(0.28 : 0.37)	0.18	(0.13 : 0.23)
α_{TR}	1.63	(1.41 : 1.86)	0.98	(0.86 : 1.10)	0.94	(0.83 : 1.06)
β_{TR}	0.56	(0.48 : 0.65)	0.50	(0.39 : 0.61)	0.46	(0.31 : 0.61)
MWIC estimates (PROM score metric) by different methods						
MWIC – LIRT (Median)[3]	4	(3 : 4)	3	(3 : 4)	3	(2 : 4)
MWIC – LIRT (Mean)	3.78	(3.31 : 4.44)	3.44	(2.61 : 4.29)	3.20	(1.90 : 4.46)
MWIC - Mean change[4]	2.74	(1.87 : 3.61)	2.59	(1.44 : 3.73)	7.81	(6.37 : 9.25)
MWIC - ROC analyses[3]	2	(0 : 2)	4	(3 : 5)	10	(8 : 10)
MWIC - Adjusted predictive model[4]	1.78	(1.40 : 2.16)	3.30	(2.86 : 3.72)	6.71	(6.17 : 7.18)

dPROM: Difference in PROM score from baseline to specified follow-up timepoint. ES: Effect size. TR: Transition Rating. LIRT: Longitudinal item response theory. θ_1: Latent score at baseline. θ_2: Latent score at specified follow-up timepoint. α_{TR}: Discrimination parameter for the transition rating item. β_{TR}: Threshold parameter for the transition rating item. MWIC: Meaningful within-individual change. [1]For model identification, θ_1 mean is fixed at 0 and θ_1 SD is fixed at 1. [2]Since $\theta_1$mean = 0, $d\theta$ mean is equal to θ_2 mean. [3]MWIC estimates based on LIRT (median approach) or ROC analyses take only whole numbers, since they are based on values observed in the data set. [4]Estimates based on mean change analyses or predictive modeling assume continuous measurement and can therefore take decimal values.

Appendix: ESTABLISHING THRESHOLDS FOR MEANINGFUL WITHIN-INDIVIDUAL CHANGE USING LONGITUDINAL ITEM RESPONSE THEORY

Statistical model

Assume that a set of items measuring a particular health construct has been administered at baseline (time 1) and follow-up (time 2). Also, a TR item covering the same construct (scored 1: meaningful improvement and 0: no meaningful improvement) has been administered at time 2. The score for person *j* on item *i* at time 1 is labeled X_{ij1}, the score at time 2 is labeled X_{ij2}. The score on the TR item is labeled TR_j. The latent health score for person *j* at time 1 is labeled θ_{j1}, the latent score at time 2 is labeled θ_{j2}, and the latent score change from baseline to follow-up is labeled $d\theta_j$.

A simple model for MWIC would be a threshold value β_{TR} indicating the minimal level of change that is deemed important by patients. So

$$TR_j^* = \rho_{TR}\, d\theta_j + \epsilon_j$$

$$TR_j^* \geq \beta_{TR} \Rightarrow TR_j = 1, \quad TR_j^* < \beta_{TR} \Rightarrow TR_j = 0 \qquad (1)$$

Where TR_j^* is a latent formulation of the TR on an underlying continuous scale. This model includes a random component ϵ_j (with a mean of 0) reflecting measurement error in the TR item. An additional contributor to ϵ_j could be between-individual differences in the threshold for important change. Assuming such between-individual differences, β_{TR} should be interpreted as the mean threshold for important change. Both measurement error and between-individual differences in thresholds for MWIC may contributed to ϵ_j; the model does not distinguish between the two contributions. ρ_{TR} is a scaling constant. If ${\rho_{TR}}^2 + var_{\epsilon_j} = 1$, then ${\rho_{TR}}^2$ can be interpreted as the reliability of the TR item.

If ϵ_j has a normal distribution, the model above implies a normal-ogive IRT model [28, 29]. If ϵ_j has a logistic distribution, the model above implies a logistic IRT model. Normal-ogive and logistic IRT models are very similar [30]. This paper uses the logistic IRT model which we have found to provide more robust results for MWIC estimation.

The logistic IRT model for the TR item can be specified in the following way:

$$P(TR_j = 1) = \frac{exp(\alpha_{TR}(d\theta_j - \beta_{TR}))}{1+exp(\alpha_{TR}(d\theta_j - \beta_{TR}))} \quad (2)$$

or, for simplicity [31]

$$ln\left(\frac{P(TR_j=1)}{P(TR_j=0)}\right) = \alpha_{TR}(d\theta_j - \beta_{TR}) \quad (3)$$

where β_{TR} is the MWIC threshold and α_{TR} is a so-called discrimination parameter reflecting the ability of the TR item to distinguish between large and small health changes. α_{TR} is inversely related to the variance of the random component ϵ_j in equation (1). For a logistic model, $\alpha_{TR} = 1.7\frac{\rho_{TR}}{\sqrt{var_{\epsilon_j}}}$ [29, 30].

To estimate the model, we need more indicators of health change. For this purpose, equation (3) can be rearranged the following way:

$$ln\left(\frac{P(TR_j=1)}{P(TR_j=0)}\right) = \alpha_{TR}(d\theta_j - \beta_{TR}) = \alpha_{TR}(\theta_{j2} - \theta_{j1} - \beta_{TR}) = \alpha_{TR}\theta_{j2} - \alpha_{TR}\theta_{j1} + L_{TR} \quad (4)$$

Here, the response on the transition rating item is modeled as a function of the latent score at baseline and follow-up. Equation 4 demonstrates the rational for imposing the restriction that the discrimination parameter for θ_{j1} is of the same magnitude but opposite sign as the discrimination parameter for θ_{j2}. The intercept parameter $L_{TR} = -\alpha_{TR}\beta_{TR}$ is often used in multidimensional IRT models. θ_{j1} and θ_{j2} can be further characterized by including IRT models for the PROM items at baseline and follow-up. Using the Graded Response IRT model [32], these models can be written as:

$$ln\left(\frac{P(X_{ij1}\geq c)}{P(X_{ij1}<c)}\right) = \alpha_i\theta_{j1} + L_{ic} \text{ for time 1, and} \quad (5)$$

$$ln\left(\frac{P(X_{ij2}\geq c)}{P(X_{ij2}<c)}\right) = \alpha_i\theta_{j2} + L_{ic} \text{ for time 2} \quad (6)$$

where α_i is the discrimination parameter for item *i* and L_{ic} is the intercept parameter for category *c* of item *i*. Item parameters (α_i and L_{ic}) are assumed to be identical for time 1 and time 2, but this assumption can be relaxed for some items. The model described by equations 4-6 can be estimated by software for multidimensional IRT, such as Mplus or the R package mirt. In this paper, the LIRT models were fitted using maximum likelihood estimation. To identify the model, the θ variance at time 1 ($var_{\theta_{j1}}$) is usually set to 1.

If $var_{d\theta_j}$=1 then $\alpha_{TR} = 1.7\frac{\rho_{TR}}{\sqrt{1-\rho_{TR}{}^2}} \Leftrightarrow \rho_{TR}{}^2 = \frac{\alpha_{TR}{}^2}{1.7^2+\alpha_{TR}{}^2}$

Thus, at follow-up 1, the α_{TR} parameter of 1.63 and $d\theta$ SD of 1.03 implies a TR item reliability of 0.49. This is in line with empirical results on the reliability of the TR item (Griffiths et al, in review).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MpluscodeLIRTforMWIC.inp
- RcodeLIRTMICinmirt.r