Toward a rigorous assessment of the statistical performances of methods to estimate the Minimal Important Difference of Patient-Reported Outcomes: A protocol for a large-scale simulation study

Antoine Vanier, Maxime Leroy, Jean-Benoit Hardouin

Version of Record: https://www.sciencedirect.com/science/article/pii/S1046202322000366
Manuscript_a1acd12cff02ed0d7c99b82400384545

Toward a rigorous assessment of the statistical performances of methods to estimate the Minimal Important Difference of Patient-Reported Outcomes: a protocol for a large-scale simulation study

Antoine Vanier[1,2], Maxime Leroy[3], Jean-Benoit Hardouin[1,3]

1 Inserm - University of Nantes - University of Tours, UMR U1246 Sphere "Methods in patient-centered outcomes and health research", Nantes 44200, France

2 Haute Autorité de Santé, Assessment and Access to Innovation Direction, Pharmaceutical Drugs Assessment Department, Saint-Denis 93210, France

2 University Hospital of Nantes, Unit of Methodology and Biostatistics, Nantes 44000, France

Corresponding author

Antoine Vanier

Inserm U1246 Sphere “Methods in patient-centered outcomes and health research”

Institut de Recherche en Santé 2 – Université de Nantes

22, Boulevard Bénoni-Goullin

44200 Nantes

E-Mail: antoine.vanier@univ-tours.fr

ABSTRACT

Interpreting observed changes over time in Patient-Reported Outcomes (PRO) measures is still considered a challenge. Indeed, concluding an observed change at group level is statistically significant does not necessarily equate this change is meaningful from the perspective of the patient. To help interpret within and/or between group changes in the measure over time, the estimation of the Minimal Important Difference (MID) of the instrument – the smallest value that patients consider as a perceived change – is useful. In the last 30 years, a plethora of methods and estimators have been proposed to derive this MID value using clinical data from sample of patients. MIDs for hundreds of PROs have been estimated, with frequently a substantial variability in the results depending on the method used. Nonetheless, a rigorous assessment of the statistical performances of numerous proposed methods for estimating MIDs by experimental design such as Monte-Carlo study has never been performed.

The purpose of this paper is to thoroughly depict a protocol for a large-scale simulation study designed to investigate the statistical performances, especially bias against a true populational value, of the common proposed estimators for MID.

This paper depicts how investigated methods and estimators were retained after the conduct of a systematic review, the design of a conceptual model that formally defines what is the true populational MID value and the translation of the conceptual model into a model allowing the simulation of responses of items to a hypothetical PRO at two times of measurement along with the response to a Patient Global Rating of Change at the second time under the constraint of a known true MID value. A statistical analysis plan is depicted in order to conclude if working hypotheses on what could be appropriate MID estimators will be verified. Strengths, assumptions, and limits of the simulation model are exposed.

Finally, we show how this protocol could be the basis for fostering future methodological research on the issue of interpreting changes in PRO measures.

Key words: Patient-Reported Outcomes, Minimal Important Difference, Minimal Clinically Important Difference, Psychometrics, Simulation study, Monte-Carlo study

I/ INTRODUCTION AND OBJECTIVES

In the field of human healthcare, *Patient-Reported Outcome*s (PROs) allow measuring quantitatively relevant *subjective constructs* such as fatigue, depression, pain, or Quality of Life [1,2]. These constructs can be assessed almost exclusively by any other means than taking the patient's perspective into account. Their measure is of paramount importance to access their thoughts, feelings, or preferences, as they go through new medical experiences such as living with a chronic disease or engaging in an intensive therapeutic course. Most of the times, PROs are self-administered questionnaires, composed of multiple *items* (i.e., questions) with a pre-specified response format (usually a Likert scale or a Visual Analogous Scale) [1]. Using a *measurement model* (i.e., an algebraic mapping), responses to items are transformed into a quantitative measure of the *latent c*onstruct of interest (i.e., latent as it is not observed directly but is supposed to explain the variability of the responses to the items) onto a *measurement scale* [3]. The most frequent measure that is used is called a *score* and is usually computed as the simple sum of codes affected to the responses to the items [4]. If a sufficient level of psychometric properties (i.e., in general, validity and reliability) are verified, the score can be taken as an appropriate ordinal measure of the construct of interest [4].

While the interpretation of the relevance of observed change in a PRO measure over time (e.g., before and after chemotherapy) is a topic of great interest in the field of healthcare psychometrics, it is still considered a challenge [3,5,6]. Indeed, first, PROs have a much shorter history of development than others historic measures in physics. Second and more importantly, due to the subjective nature of the targeted constructs, the calibration of scales is relative to internal standards in people's mind (e.g., when assessing pain on a single Visual Analogous Scale item, it is the patient who defines what is "absence of pain", what is "the most pain I have ever experienced", and what change in the "true" level of pain constitutes a change of one unit in the measure) [7]. Therefore, the interpretation of a change on such a scale is often considered difficult by clinician and researchers. When, for example, an average increase of 4 points in Quality of Life on a scale from 0 to 100 is observed on a group of patients over time, a rejection of the null hypothesis of no change using a test statistic is not sufficient to assume this change is meaningful for the patients [8].

To enhance the interpretability of observed change in PRO scores over time, the search for a relevant *threshold* allowing the partition of patients as having experienced a meaningful change in PRO scores or not is a frequent strategy [9]. This threshold has been

defined by the US Food and Drug Administration since 2009 as the *responder definition*: "*the individual Patient-Reported Outcomes (PRO) score change over a predetermined time period that should be interpreted as a treatment benefit*" [10]. This search for a responder definition value can be obtained through a choice of a numerous perspectives (e.g., from the point of view of a healthcare professional, or from the point of view of healthcare policy makers) [11]. The most frequent perspective is to estimate a threshold that has meaning according to the patient's perspective. When doing so, this threshold is usually called the *Minimal Clinically Important Difference* (MCID) or *Minimal Important Difference* (MID). This notion was firstly defined by Jaeschke et al. in 1989 as follows: "*the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management*" [12]. Since 1989, estimating the MID of a given PRO has been the subject of hundreds of studies on empirical clinical data [13,14].

Most methods to estimate a MID belong to two categories: *anchor-based methods* and *distribution-based methods* [5,13,15,16]. Anchor-based methods link the observed change in scores to an external indicator (*the anchor*) classifying patients as improved, worsened, or stable. The most used anchor is a *Patient Global Rating of Change* (PGRC) [14]. It is a single item used at the second time of measurement (e.g., after a surgery) assessing an overall feeling of perceived change since the baseline assessment [12]. An example would be: "*Compared to before your surgery, overall, do you think your quality of life is now…*". Response options can be "*a lot worse*", "*a little worse*", "*about the same*", "*a little better*", and "*a lot better*". This assessment of perceived change is then used to estimate a MID threshold. An example would be to estimate a MID as the mean of observed change in scores within the subgroup of patients who have experienced a little change according to the PGRC [15]. As anchor-based methods involve an explicit assessment of perceived change by the patient, they are frequently considered appropriate to estimate a responder definition threshold according to the patient perspective, despite being prone to recall bias [8,17]. On the contrary, distribution-based methods use exclusively the variability of PRO scores (either the score at baseline or the observed change in scores) as data to derive a MID [15,16]. They can be based on Cohen's *effect sizes* [18]. An effect size is obtained by dividing the mean change in scores by the standard deviation of the baseline score. For example, based on data from studies in experimental psychology, Cohen has proposed that *0.5 effect size* corresponds to a change "*visible by the naked-eye*". This value is frequently used as a MID estimate [14,19].

Distribution-based methods can also be based on discriminating an observed change as true signal from measurement error [15,16]. Based on empirical observations, *1 Standard Error of Measurement* of the score at baseline is also frequently assumed to be a plausible estimate of the MID [14,20,21].

Numerous methods and estimators have been proposed for deriving a MID value [13–16]. However, to this date, there is no consensus on the appropriate method(s) to use. Little to nothing is known about the statistical performances of almost all estimators, especially *bias* against a true populational MID value. Therefore, some authors recommend to "*triangulate*" the MID of a given PRO by using multiple estimators from different family of methods to come up with a plausible range within which the true MID value is [22]. Nonetheless, the application of any statistical methods to empirical data, whether triangulated or not, does not allow studying their statistical performances, such as bias, as the "truth" behind the generation of the data is unknown.

Monte-Carlo simulation studies are appropriate experimental designs to investigate the statistical performances of methods. They are "*computer experiments* (or *in silico* designs) *that involve creating data by pseudo-random sampling from known probability distributions*" [23]. Simulation studies must be used when the performances of a statistical technique cannot be assessed analytically because of complexity (e.g., a simulation study would be useless to study the sampling variability of a sample mean because theories like the Central Limit Theorem provide the necessary solution analytically). As *true values* of statistical parameters used to simulate data are known and therefore can serve as reference, simulation studies can be used to assess the statistical performances of methods such as bias under ideal circumstances or robustness to varying sample characteristics (e.g., sample size…) [23]. Simulation studies can be used to assess the statistical performances of a single method, but comparisons of the relative performances of multiple methods in estimating the same populational parameter can also be performed. In the realm of modeling longitudinal PROs data, they have been frequently used to assess the robustness of certain models to the presence of missing data in terms of bias, or power and control of type-I-error of complex algorithms in detecting phenomenon such as lack of measurement invariance [24–27]. Two aspects can be considered critical when designing a simulation study. First, as they are based on creating artificial data, a simulation model generating the data must be designed. While the model is almost certainly a simplification of the "true" data generation mechanism, it needs to be plausible (i.e., it needs to mimic a theory of data generation with sufficient plausibility).

Second, when conducting a simulation study, many methods can be assessed under many scenarios (e.g., variations in sample size or amount of missing data) and in each scenario pseudo-random sampling is replicated many times (to obtain sufficient precision in estimating statistics of interest such as bias). Therefore, these studies can produce a huge amount of data and proposing an appropriate statistical analysis plan can be challenging. Although these studies can have a high level of complexity and are built under strong assumptions, the methodology of a Monte-Carlo study is frequently under-reported in papers investigating methods for analyzing longitudinal PRO data [28].

To our knowledge, we are currently performing, for the first time, a large-scale simulation study investigating the statistical performances of common MID estimators found in the literature under various scenarios, by simulating responses to the items of hypothetical PROs and to a PGRC at two times of measurement under the constraint of a known true MID value. To do so, we have faced numerous conceptual and methodological challenges and our simulation model is built under specific assumptions. Therefore, the objective of this paper is to propose a comprehensive protocol describing the details and assumptions of our experimental design.

First, we will briefly describe how we have selected the methods for estimating a MID we have investigated in this study and what are their characteristics. Second, we will summarize an overarching *conceptual model* we have built [29]. This model is a proposed theory describing the relationships between the different agents that are engaged when people answer to a PRO at two times of measurement and to a PGRC at the second time. This conceptual model allows us to formally define the MID as a statistical parameter in the population. Third, from the conceptual model, we will deduce working hypotheses on the plausible appropriate estimators of a MID. Fourth, we will describe in detail the different steps of the *simulation model* and its assumptions. We will also describe the scenarios we have simulated. Fifth, we will propose a statistical analysis plan for some of the investigated scenarios to confirm or not our hypotheses. Last, we will discuss the implications of the assumptions and the limits of our simulation study.

II/ PREREQUSITES

1. Methods for estimating MIDs retained in this study

A systematic literature review was conducted from February 2017 to January 2018 on articles published between January 1989 (the year the MCID was defined) and February 2017 [14]. Included articles were any study about a PRO (according to the Food and Drug Administration definition [10]) and in which there is a report of at least one empirical estimate of a MID, published in English or French. We searched potential articles to be included using MEDLINE and PsychInfo databases. A search for references within reviews of the MID concept was also performed and has helped to identify the first papers published in the early 1990s. Because we wanted to capture an a-priori unknown number of methods, the search equations (MesH terms, free text terms and synonyms) were designed to be of high sensitivity.

In the end, 328 articles were included, which corresponds to 474 PROs assessed (269 unique). The characteristics of 945 MID estimates were collected and a qualitative and quantitative analysis was performed to identify and classify the methods for MID estimation that we found. First, as the distinction between anchor-based and discrimination-based methods is consensual, this first layer of classification was used for each estimator we have found. Then, within these two main categories, we have classified each estimator into subtypes based on their proximity in terms of statistical modeling to derive an MID value. For example, all distribution-based estimators based on the idea of using Cohen's effect sizes were classified into the same subtype. Then, for anchor-based methods, additional characteristics were collected and taken into considerations for designing the simulation study. These characteristics are the number of response categories to the PGRC, the way the answers to the PGRC are used to define the group that is considered to have experience a "little change" or as having experienced "a change", the way the answers to the PGRC are used to define the group that is considered to have been stable. These additional characteristics can modify the way certain subtypes (or sometimes all subtypes) are used to derive a MID value, thus increasing the number of methods for MID estimation. These characteristics, as well as the subtypes for which they apply, are described further in the manuscript when appropriate and are summarized in Table 1. For the rest of the manuscript, we will make the distinction between a MID estimator (i.e., the subtype) and a method for MID estimation (the whole process for deriving a MID value). It would not be manageable to fully describe each method within the manuscript. A comprehensive description of each method retained in this study is available as a supplementary material (eText1).

Regarding anchor-based methods, 31 different estimators were found. We have classified them into five subtypes: 1. the use of the *mean of the "little change" group*; 2. the search for a discriminative threshold between the " change"' group and the "stable group" using either *Receiver-Operating Curve* (ROC) analysis, *predictive modeling* by logistic regression, or discriminant analysis; 3. the mean of the "little change" group modeled by linear regression, 4. the use of the *75th percentile* of the "little change" group, and 5. other various estimators. Then 21 distribution-based estimators were found. We have classified them into four subtypes: 1. those based on effect sizes, 2. those based on *Standardized Response Means*, 3. those based on measurement error, and 4. those based on the range of the scale of change in scores.

We chose not to retain all estimators that were found. First, we have excluded estimators that cannot be assessed within our simulation framework (i.e., estimators using data from more than two times of measurement). Second, we have decided to exclude the estimators based on the mean of the "little change" group modeled by linear regression as we have considered these estimators as equivalent to the MID as the sample mean estimate of the "little change" group. Third, we have excluded estimators that were used in less than 1% the studies we have assessed. Last, an exception was made for the estimator based on predictive modeling by logistic regression. While this estimator was used in less than 1% of the studies we have assessed, this was the only estimator for which simulations studies suggesting interesting statistical properties were performed [30–32].

In the end, we have retained five main anchor-based estimators:

1. the MID as the mean of the change in scores within the "little change" group according to the PGRC [12],

2. the MID as the mean of the change within the "little change" group *minus* the mean of the change in scores within the stable group [33],

3. the MID as the 75th percentile of the change in scores within the "little change" group [34],

4 the MID as a threshold in the change in scores discriminating patients who are classified as having changed from those classified as not according to the PGRC using a ROC analysis [15],

5. the MID as a threshold using predictive modeling where the threshold is algebraically derived from the estimates of a logistic regression with the classification of

patients as changed or not as the dependent variable and the change in scores as the independent variable [30].

Applying to all the anchor-based methods, two other design characteristics are explored in this simulation study. First, we found it is frequent to estimate a MID for improvement (i.e., for patients who feel they are better at the second time of measurement) and a MID for deterioration (i.e., for patients who feel they are worse at the second time of measurement). This study explores both cases. Second, the number of response categories to the PGRC can vary: it can be five ("a lot worse", "a little worse", "stable", "a little better" "a lot better"), seven ("a lot worse", "somewhat worse", "a little worse", "stable", "a little better", "somewhat better", "a lot better") or three ("worse", "stable", "better") [14].

Anchor-based methods using ROC analysis or predictive modeling are two estimators grounded in the idea of discriminating two populations with the best threshold value possible (i.e., the MID estimate). They require classifying some or all the patients into two groups based on the answers to the PGRC: 1. those classified as having experienced a change and 2 those classified as not. We explore four different ways of using the responses to the PGRC to classify the patients into two groups (changed or not) (Figure 1): 1. "little change" patients only *versus* stable patients only, 2. "little change" patients only *versus* the rest of the patients, 3. "little + somewhat change" patients *versus* stable patients (PGRC with 7 categories only), 4. "little + somewhat change" patients *versus* the rest of the patients (PGRC with 7 categories only) [14,35].

For anchor-based methods based on ROC analysis, a statistical criterion for discriminating the patients that have changed from those that have not has to be used. The choice of the criterion can have an impact on the MID estimate. We explore five different criteria for finding the MID threshold:

1. the point the closest to the top-left of the cartesian plan in Euclidian distance,
2. the point maximizing the Youden index,
3. the point allowing obtaining at least a specificity of 80%,
4. the point allowing obtaining a specificity of 100%,
5. the point allowing obtaining a sensitivity of 100% [14].

For the predictive modeling estimators, two algebraic ways of deriving the MID estimate were proposed: 1. a crude one and 2. an algebraically adjusted one on the observed proportion of patients who have changed according to the PGRC [30,31]. The study explores both estimates.

Finally, we have retained eleven distribution-based estimators. Three are based on effect sizes: the MID as 0.2 effect size, 0.5 effect size, or 0.8 effect size. Three are based on standardized response means: the MID as 0.2 standardized response mean, 0.5 standardized response mean, or 0.8 standardized response mean. Three are based on the measurement error: the MID as 1 Standard Error of Measurement, 1.96 Standard Error of Measurement or 1 Minimal Detectable Change [15,16]. Two are based on the range of the scale of the change in scores: 8% of the range of the change and 7% of the total change possible [14]. These estimates do not distinguish the MID for improvement from the MID for deterioration.

In all, the study investigates 70 MID methods for MID estimation. Their classification is summarized in Table 1.

2. A conceptual model defining the "true MID" as a populational value

Since the inception of the MCID notion in 1989, empirical estimates of this parameter were performed on hundreds of clinical datasets [13,14]. Nonetheless, the MID was never formally defined. The lack of definition of a true MID as a populational parameter was a major caveat for our simulation study. Indeed, to investigate the bias of the proposed methods for estimating MIDs, it is necessary to simulate data under the constraint of a known populational MID value which is the "truth" against which sample estimates are compared. Therefore, before planning the simulation study, we have proposed a conceptual model describing the relationships between the agents engaged when patients answer to PROs items at two times of measurement and to a PGRC at the second time of measurement [29]. From this model, we were able to propose a formal definition of the true MID value. This proposal was fully developed in a previous paper [29]. The development of this conceptual model was based on an adaptation of an already existing model: the Rapkin and Schwartz model, published in 2004 [36]. This model describes the agents engaged in explaining change in Health-Related Quality of Life over time, using concepts from the field of psychology of survey response [37]. Briefly, first, we have adapted this model such as it describes the agents engaged in explaining the level of any subjective construct at two times of measurement. Then, we have postulated the necessary occurrence of two additional cognitive constructs (the perceived change and the remembered baseline level of the latent trait (which will be described further)) to explain how someone answers to a PGRC at a second time of measurement, and we have delineated their plausible relationships with the other agents. Plausible paths between the different agents were proposed based on the literature. Finally,

from the model, we were able to propose a plausible definition of the MID at the populational level.

As mentioned in the introduction section, a responder definition threshold can be estimated from several other perspectives than the patient perspective. In addition, the concept of the MID has generated various debates. Some of these issues involve the distinction about getting a threshold that characterizes a change as "minimal" *versus* "meaningful" or the non-ambiguous meaning of "important" [38]. Thus, to be specific, we need to clarify what perspective we adopt. The responder definition threshold we define corresponds to the minimal amount of change in PRO scores that is subjectively considered a change by the patient. We call it the "*Minimal Perceived Change*" (MPC) and its definition is "*the minimum amount of change in PRO scores over time that is perceived by a person as a nonstable trajectory*" [29]. Therefore, for the rest of the manuscript, we will use the term MPC instead of MID.

The full conceptual model is presented in Figure 2. A construct of interest (e.g., Quality of Life, fatigue…) is measured at two times (SC_{t1} and SC_{t2}). To answer the PRO items at each time of measurement, personal appraisal processes are elicited. These are the cognitive processes that are needed to select the desired answers to the items (A_2 and A_6 bidirectional paths). Between the two measurements, there is the occurrence of a catalyst: one or multiple event(s) or life experience(s) susceptible to trigger a change in the target construct (e.g., the diagnosis of a cancer susceptible to decrease Quality of Life) (S_4 path) [39]. The catalyst can trigger psychological mechanisms to buffer the effect of the catalyst on the target construct (C_3 then C_4 paths). Antecedents are more or less stable personal or environmental characteristics (e.g., personality, socioeconomic status) which set the baseline conditions. Answering the PGRC at the second time of measurement is a process that appraised a cognitive construct we call the perceived change: an overall feeling of change conceived as continuous. To elicit this feeling of Perceived Change in someone's mind, there is a need to remember what the level of the target construct at baseline was (SC_{t1mem}). The P_2 path represents the ability of correctly remembering what the baseline level was. But SCt_{1mem} can also be reconstructed from the present state (P_5 path). Antecedents can explain the ability to remember (P_1 path) as well as the catalyst and psychological mechanisms (e.g., a head trauma resulting in memory impairment) (P_3 and P_4 paths). Finally, the level of perceived change is set as the difference between SC_{t2} and SC_{t1mem} (P_7 minus P_6 paths).

Then, selecting a response category to answer a PGRC is akin to discretize a continuous state (i.e., the level of perceived change) into one of the proposed answers. To achieve this, one needs to set several thresholds of perceived change values defining the bounds for switching from one category to the next (e.g., from "the same" to "a little better"). We denote these PGRC thresholds T_s, and for a PGRC with *K* categories, there are *K-1* T_s. By comparing the level of perceived change with adjacent T_s, one can select the desired answer to the PGRC. As appraisal processes are psychological personal characteristics, we assume each T_s is a random variable with a distribution in the population (i.e., calibrating the value of the thresholds on the perceived change scale is relative to internal standards). If T_1 and T_{-1} are the perceived change values for switching from "the same" to "a little change" (either improvement or deterioration), they therefore have a distribution in the population with a location and dispersion parameter: $T_1 \sim D(\lambda_1, \zeta_1)$ and $T_{-1} \sim D(\lambda_{-1}, \zeta_{-1})$. The MPC for improvement is therefore $MPC_+ = \lambda_1$ and $MPC_- = \lambda_{-1}$ (Figure 3).

III/ WORKING HYPOTHESES

Because we have formally defined what is the true population MPC value, we are also able to deduce hypotheses about what could be appropriate estimators of the MPC (these hypotheses were already partly exposed in the paper about the conceptual model) [29].

First, among anchor-based methods, we hypothesize that appropriate estimators which indeed target the populational MPC value are the ones based on discriminating the population of unchanged patients from whose who have changed according to the response of the PGRC. Thus, we hypothesize than estimators based on ROC analysis or predictive modeling will be unbiased estimators of the populational MPC, at least under the assumption of a perfect recollection of the SC_{t1} value at the second time of measurement. In addition, we hypothesize distribution-based methods, as they do not use the data obtained from the response to the PGRC, can target the populational MPC value by chance only and thus will all be biased in this simulation study when averaging the results on all the explored scenarios.

Three simulations studies from the same group of researchers have been published about the statistical performances of predictive modeling against ROC analysis (using the Youden index) [30–32]. If there are major differences between their simulation model and ours (it is discussed below in the paper), we do think some of their conclusions can be the basis of specific hypotheses our study can confirm or not. Two main results of their studies

are of interest and can be investigated in our study. First, under ideal circumstances (perfect recollection of the value of SC_{t1} and a proportion of improved or worsened patients equal or close to 50% according to the response of the PGRC), estimators based on predictive modeling or ROC analysis are unbiased, but those based on predictive modeling exhibit a better precision. Then, when the proportion of improved or worsened patients is not equal or close to 50%, the adjusted estimator based on predictive modeling will be the only one unbiased.

IV/ SIMULATING SAMPLES OF PATIENTS

1. Overview

The simulation model allows us to simulate responses to items at two times of measurement and a response to a PGRC at the second time. The model is an operationalization in variables and mathematical functions of the conceptual model. The aim is to obtain simulated empirical samples of people drawn from a population with known true parameters while respecting a trade-off between simulating plausible conditions and ease of simulation.

To simulate all the necessary data for a given dataset, three main steps are required (Figure 4):

1. the simulation of responses to items of a PRO questionnaire at two times of measurement (red and orange boxes of Figure 4). Each time, the responses are supposed to be caused by the level of a unique latent trait ($\theta^{(1)}$ and $\theta^{(2)}$) (i.e., hypothesis of unidimensionality). The latent trait operationalizes the hypothetical construct of interest (i.e., SC_{t1} and SC_{t2}; e.g., Quality of Life, fatigue, pain, depression, anxiety). It has a distribution in the population with a mean and variance. The level of the latent trait is correlated at the populational level between the two measurement occasions;
2. the simulation of necessary variables and structural relationships to estimate for each person an individual level of SCt_{1mem} (the remembered baseline level of the latent trait) and perceived change (blue box of Figure 4);
3. the simulation of a response to a polytomous PGRC at the second time of measurement. It is obtained by a process of discretization of the individual level of

perceived change (conceived as continuous) onto one discrete state (i.e., the response to the PGRC) (green box of Figure 4).

2. Generating responses to items at two times of measurement

The purpose of this first step is to simulate plausible individual polytomous responses to items at two times of measurement for each person of a hypothetical dataset under two constraints: a known distribution in the population of the latent traits (i.e., the operationalization of the construct of interest as variables) at both times of measurement, and the simulation of responses to item of an hypothetical PRO questionnaire with an adequate level of reliability and structural validity (i.e., assuming the hypothesis of unidimensionality).

To do so, a measurement model such as a longitudinal Partial Credit Model (lPCM) from Rasch Measurement Theory (RMT) can be considered adequate [25]. The choice of a model coming from the RMT allows satisfying the fact that the simple sum score of the code affected to the responses to the items is an adequate ordinal measure of the concept to measure (latent trait). Indeed, models coming from the Rasch Measurement Theory verify the property of the sufficiency of the score on the latent trait : for each value of the score, there is only one possible estimation of the latent trait [40].

Let simulate the responses of *N* patients to *J* polytomous items. We assume that the items have *M* positive response categories, and these measures are repeated 2 times on the N patients of the study. The response of patient *i (i = 1;…;N)* to an item *j (j = 1;…; J)* at time *t (t = 1;2)* is denoted by $X_{ij}^{(t)}$.

The measurement model is as follows:

$$P\left(X_{ij}^{(t)} = h|\theta_i^{(t)}, \delta_{j1}, \dots, \delta_{jm_j}\right) = \frac{exp\left(h\theta_i^{(t)} - \sum_{l=1}^{h(t)} \delta_{jl}\right)}{\sum_{c=0}^{m_j} exp\left(c\theta_i^{(t)} - \sum_{l=1}^{c} \delta_{jl}\right)} \textit{(Eq 1).}$$

With:

h: a specific possible answer to item j,

$\theta_i^{(t)}$: level of latent trait of patient i at time t,

$\left(\delta_{jl}\right)$: item threshold of item j (l = 1;…;m_j).

It models the probability for an individual i to answer h to item j at time t as a function of a person characteristic (its level of latent trait $\theta_i^{(t)}$) and a characteristic of the items called item threshold parameters $\left(\delta_{jl}\right)$. For an item j with m+1 response categories (modalities),

there are m_j item threshold parameters. When an item has high values of item thresholds, a high level of latent trait is necessary to have a high probability of answering the highest modality of the item (e.g., if the latent trait is physical functioning, a corresponding item could be "How is it easy for me to run a marathon": a very high level of physical functioning is required to have a high probability to answer the highest category ("Very easy")). Contrarywise, an item with a low level of item thresholds could be "How is it easy for me to run 20 meters".

The latent trait is distributed as follow: $\theta \sim N\left(\begin{pmatrix}\alpha^{(1)}=0\\ \alpha^{(2)}\end{pmatrix}, \begin{pmatrix}\sigma^{2(1)}=1 & \rho\\ \rho & \sigma^{2(2)}\end{pmatrix}\right)$ with $\alpha^{(t)}$ the mean value of the latent trait at time t, $\sigma^{2(t)}$ the variance at time t, and ρ the between-time covariance.

The matrix of items thresholds (δ_{jl}) is chosen to reflect a well-adapted questionnaire for the population: the global item distribution is centered on $\alpha^{(1)} = 0$ at time 1 and the respective items thresholds are regularly spaced to cover the whole distribution of the latent trait while avoiding floor and ceiling effects. A graphical illustration for 5 dichotomous items and a distribution of the latent trait $\theta^{(1)} \sim N(\alpha^{(1)} = 0, \sigma^2_{\theta(1)} = 1)$ is proposed as Figure 5.

Therefore, the random draw of items thresholds is as such:

- $\forall j\ \delta^*_{j1}$ are the $j/(J+1)^{th}$ percentiles of $N\sim(0,1)$ if $\theta^{(1)} \sim N(\alpha^{(1)} = 0, \sigma^2_{\theta(1)} = 1)$,
- $\forall j\ \delta^*_{jl} = \delta^*_{j1} + 2(l-1)/(m_j - l)$ for $1 < l \leq m_j$,
- $\bar{\delta} = \frac{\Sigma_{j=1}^{j}\Sigma_{l=1}^{m_j}\delta^*_{jl}}{\Sigma_{j=1}^{j} m_j}$,
- $\forall j, l\ \delta_{jl} = \delta^*_{jl} - \bar{\delta}$.

An example with J = 5 and $\forall j, m_j = 3$ would be:

$$(\delta_{jl}) = \begin{pmatrix}-1.962 & -0.962 & -0.038\\ -1.432 & -0.432 & 0.568\\ -1.002 & -0.002 & 0.998\\ -0.572 & 0.428 & 1.428\\ -0.032 & 0.968 & 1.968\end{pmatrix}.$$

We assume people appraise the items at time 2 the same way than time 1. Thus, the items thresholds remain constant from time to time (i.e., longitudinal measurement invariance is assumed).

Simulated values for each person of correlated latent traits level at time 1 and time 2 are as follow:

- for each patient, i, values are randomly drawn for two independent variables Z_1 and Z_2: $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$,
- at time 1: $\theta_i^{*(1)} = Z_{1i}$,
- at time 2: $\theta_i^{*(2)} = \rho Z_{1i} + \sqrt{1-\rho^2} Z_{2i}$,,
- it follows: $\theta^* \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$.

These "standardized" latent traits are then transformed to obtain individual latent traits values drawn from the desired distribution at time 1 and 2:

- $\theta_i^{(1)} = \theta_i^{*(1)} \times \sigma_{\theta(1)} + \alpha^{(1)}$, $\theta^{(1)} \sim N(\alpha^{(1)}, \sigma^2_{\theta(1)})$,
- $\theta_i^{(2)} = \theta_i^{*(2)} \times \sigma_{\theta(2)} + \alpha^{(2)}$, $\theta^{(2)} \sim N(\alpha^{(2)}, \sigma^2_{\theta(2)})$,
- it follows $cor(\theta^{(1)}, \theta^{(2)}) = \rho$.

Finally, for each person, the responses to the items are simulated as follow:

- $p_{ijht} = P\left(X_{ij}^{(t)} = h^{(t)} | \theta_i^{(t)}, \delta_{j1}, \dots, \delta_{jm_j}\right)$ (calculated using *Eq 1*),
- $uni \sim U(0,1)$,
- $\mathrm{X}_{\mathrm{ij}}^{(\mathrm{t})} = \mathrm{h}$ if $\sum_0^{\mathrm{h}-1} \mathrm{p}_{\mathrm{ijht}} < \mathrm{uni} \leq \sum_0^{\mathrm{h}} \mathrm{p}_{\mathrm{ijht}}$.

This algorithm to simulate the responses to the items permits a stochastic attribution of values and therefore simulate the occurrence of measurement error.

3. Generating the individual values of the perceived change

For simulating individual values of perceived change, the simulation of individual values of SC_{t1mem} (the remembered baseline level of the target construct) is first required. SC_{t1mem} is conceptualized as a function of the level of the target construct at time 1 and time 2, and as a function of other contingencies (antecedents, catalyst, mechanisms (Figure 2 and Figure 4)). As at least antecedents and mechanisms at time 2 can vary between individuals, it makes sense to represent in the model the effect of the other contingencies as a random variable with different values for everyone and not just by a fixed "disturbance coefficient".

For the sake of simplicity, the remembered baseline level of the latent trait is modelled as a linear function of three variables:

$$\theta_i^{r(2)} = \beta_1\theta_i^{(1)} + \beta_2\theta_i^{(2)} + \beta_3 U_i \textit{ (Eq 2).}$$

With:

$\theta_i^{r(2)}$: level of remembered baseline level of the latent trait of patient i at time 2,

U_i, $U \sim N(\mu_u = 0, \sigma_u^2 = 1)$: level of the random variable pooling the specific influences of other contingencies (i.e., antecedents, mechanisms, possibly the catalyst) not already explained by the value of $\theta_i^{(1)}$ and $\theta_i^{(2)}$.

$\beta_1, \beta_2, \beta_3$: the relative importance of each of the three variables on determining θ_i^r.

By constraint: $\beta_1 + \beta_2 + \beta_3 = 1$.

As $\theta^{r(2)}$ is a linear combination of $\theta^{(1)}$, $\theta^{(2)}$ and U, each following a normal distribution, it follows: $\theta^{r(2)} \sim N(\mu_{\theta^{r(2)}} = \beta_2\alpha^{(2)}, \sigma^2_{\theta^{r(2)}} = \beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2\rho)$, (see analytical proof in Supplementary eText2).

By varying the value of β_1 and β_2, it is possible to simulate different situations. For example, $\beta_1 = 1$ corresponds to a situation where the latent trait level at baseline can be remembered perfectly.

Finally, the individual perceived change level is determined as follows (Figure 4):

$$\theta_i^{pc(2)} = \theta_i^{(2)} - \theta_i^{r(2)} \textit{(Eq 3).}$$

With $\theta_i^{pc(2)}$: level of perceived change of patient i at time 2.

It follows: $\theta^{pc(2)} \sim N(\mu_{\theta^{pc(2)}} = \alpha^{(2)}(1 - \beta_2), \sigma^2_{\theta^{pc(2)}} = 1 + \beta_1^2 + \beta_2^2 + \beta_3^2 + 2(\beta_1\beta_2\rho - \beta_1\rho - \beta_2))$ *(Eq 4)* (see analytical proof in Supplementary eText2). Supplementary eTable1 shows the values of the true correlation in the population between the true change ($\theta^{(2)} - \theta^{(1)}$) and the perceived change ($\theta^{pc(2)}$) as a function of different values of β_1, β_2, and β_3.

4. Generating the responses to a PGRC under the constraint of a known populational MPC value

A response to a categorical PGRC is akin to truncate a latent continuous indicator (i.e., the perceived change ($\theta^{pc(2)}$)) following a normal distribution. Thus, for a PGRC with K categories of responses, there are K-1 PGRC thresholds (with K an uneven natural number at least equal to 3). These thresholds are values defining how the categories of the manifest ordinal indicator (i.e., the PGRC) and the latent continuous indicator (i.e., the perceived change) are related. These PGRC thresholds are the values at which there is a switch from one category to another of the observed categorical indicator. We assume these PGRC thresholds are random variables: their values vary between individuals in the population of interest.

Let simulate the responses of N patients to a PGRC at time 2 (*Y* with K categories of possible responses). This response is denoted $Y_i^{(2)}$. For a PGRC with K categories, there are K-1 PGRC thresholds (T_s) with $s = \{-1,1\}$ if $K = 3$, $s = \{-2,-1,1,2\}$ if $K = 5$ and $s = \{-3,-2,-1,1,2,3\}$ if $K = 7$ defining the symmetrical position of the threshold from 0 (the hypothetical point of no perceived change at all).

Each PGRC threshold is a random variable with a distribution in the population. We assume the shape of a Gaussian distribution, thus: $T_s \sim N(\lambda_s, \zeta_s^2)$.

To set the distribution of the PGRC thresholds, we can scale them as a function of the number of response categories of the PGRC (K), and on the standard deviation of the perceived change ($\theta^{pc(2)}$) as follows:

$$T_s \sim N\left(\lambda_s = Sign(s)\frac{2\sigma_{\theta^{pc(2)}}(2|s|-1)}{K-1}, \zeta_s^2 = \left(\frac{D}{K-1}\right)^2\right) \text{ (Eq 5).}$$

With $Sign(s) = 1$ if $s > 0$, $Sign(s) = -1$ if $s < 0$, and *D* a dispersion factor which can take a real value.

With this operationalization of the mean and variance of the PGRC thresholds in the population, we assume:

- a symmetry of the thresholds around 0 (the point of no perceived change at all),
- an equal interval between each of the thresholds on the same side of the distribution of the perceived change from 0,
- the distributions of the thresholds are a function of the number of response categories K to the PGRC, with K an uneven natural number at least equal to 3,
- the variance of each threshold is the same (for a given k).

We can note:

- $T_1 \sim N(\lambda_1 = \frac{2\sigma_{\theta^{pc(2)}}}{K-1}, \zeta_1^2 = \left(\frac{d}{K-1}\right)^2)$ *(Eq 6.1)*,
- and $T_{-1} \sim N(\lambda_{-1} = -\frac{2\sigma_{\theta^{pc(2)}}}{K-1}, \zeta_1^2 = \left(\frac{d}{K-1}\right)^2)$ *(Eq 6.2)*.

It follows:

- $MPC_+ = \lambda_1 = \frac{2\sigma_{\theta^{pc(2)}}}{K-1}$ (the true MPC value for improvement in the population),
- $MPC_- = \lambda_{-1} = -\frac{2\sigma_{\theta^{pc(2)}}}{K-1}$ (the true MPC value for deterioration in the population).

Finally, for a patient i, the response $Y_i^{(2)}$ to the PGRC at time 2 is determined by:

$$Y_i^{(2)} = \begin{pmatrix} if\ \tau_{-1,i} < \theta_i^{pc(2)} \leq \tau_{1,i} \rightarrow 0 \\ else\ if\ \theta_i^{pc(2)} \leq \tau_{min(s),i} \rightarrow -\frac{K-1}{2} \\ else\ if\ \tau_{max(s),i} \leq \theta_i^{pc(2)} \rightarrow \frac{K-1}{2}. \\ else\ if\ \tau_{s-1} < \theta_i^{pc(2)} \leq \tau_s \rightarrow s \end{pmatrix} \textit{(Eq 7)}.$$

With $\tau_{s,i}$ the value of the PGRC threshold s of an individual i.

A comprehensive formal definition of random variables and parameters of the study is proposed in Supplementary eText2.

5. Simulation parameters and explored scenarios

In this study, several person parameters or PRO characteristics have unique values for all the scenarios that are explored, as either these characteristics are constrained fixed values, or it seems plausible they do not explain a potential variation in the statistical performances of MPC estimation. Contrarywise, there are other parameters which will vary from scenarios to scenarios, as it seems plausible they can explain bias in MPC estimation. For these parameters, we have selected either values that can be encountered when analyzing clinical data, or values for exploring specific situations.

Person parameters with unique values for all the scenarios are:

- the distribution of the latent trait level at time 1: $\theta^{(1)} \sim \mathrm{N}(\alpha^{(1)} = 0, \sigma^2_{\theta(1)} = 1)$,

- the variance of the latent trait level at time 2: $\sigma^2_{\theta(2)} = 1$,
- the correlation between the two latent traits: $\mathrm{cor}(\theta^{(1)}, \theta^{(2)}) = \rho = 0.7$,
- the distribution of the variable representing the influences of several contingencies on remembering the baseline latent trait level at time 2: $U \sim N(\mu_u = 0, \sigma^2_u = 1)$.

PRO characteristics with unique values for all the scenarios are:

- the number of response categories to all the items: $M = 4$,
- the matrix of items thresholds at each time of measurement (δ_{jl}) (see values for each possible number of items in Supplementary eText2).

Person or PROs characteristics or parameters that are explored from scenarios to scenarios are (Table 2):

- the sample size: $N = \{200, 500, 1000\}$,
- the true change in the latent trait over time: $\alpha^{(2)} = \{-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8\}$,
- the importance of $\theta^{(1)}$ on determining $\theta^{r(2)}$: $\beta_1 = \{0, 0.1, 0.2, 0.3, 0.5, 1\}$,
- the importance of $\theta^{(2)}$ on determining $\theta^{r(2)}$: $\beta_2 = \{0, 0.1, 0.2, 0.3, 0.5\}$,
- the variance of the PGRC thresholds: $\zeta^2_s = \left\{\left(\frac{1}{K-1}\right)^2, \left(\frac{0.5}{K-1}\right)^2\right\}$ corresponding to D = 1 or D = 0.5,
- the number of items: $J = \{5, 10, 20\}$,
- the number of response categories to the PGRC: $K = \{3, 5, 7\}$.

The two different values of the variance of the PGRC thresholds (ζ^2_s) were chosen to reflect a situation where thresholds are quite heterogeneous within the population or more homogeneous. The choices of 1 and 0.5 as the dispersion factor were also determined for ensuring a low probability of inadequate ordering of the PGRC thresholds along the perceived change continuum. If it happens during the simulation process, it is prevented by the following conditional statement:

$$\begin{pmatrix} if\ \tau_{\min(s),i} < \tau_{\min(s)+1,i} < \cdots < \tau_{\max(s)-1,i} < \tau_{\max(s),i} = FALSE \rightarrow\ I_{(i)} = 0 \\ while\ I_{(i)} = 0 \rightarrow\ draw\ new\ \tau_{s,i}\ until\ \ I_{(i)} = 1 \end{pmatrix}.$$

With $I_{(i)}$ an indicator variable taking the value 0 or 1 for a patient i.

For illustration, Figure 6 shows the distribution of thresholds on the scale of the perceived change for K = 3, or 5; $\alpha^{(2)} = 0$ and assuming $\theta^{r(2)}$ is a perfect recollection of $\theta^{(1)}$ (thus, $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = 0$, in that case $\theta^{pc(2)} \sim N(0, 0.6)$), for $\zeta_s^2 = \left(\frac{1}{K-1}\right)^2$.

In the end, seven person or PRO characteristics can vary from scenarios to scenarios (Table 2). The combination of all these values leads to 9 828 explored scenarios. For each scenario, 500 sample replicates are simulated.

V/ STATISTICAL ANALYSIS PLAN

1. Prerequisites

After the simulation of all the 500 sample replicates for the 9 828 explored scenarios, observed scores (i.e., the sum of the response to the items of the simulated PROs) are estimated at each time of measurement for all the individuals of all simulated datasets. This allows using the responses to the items as a basis for measuring the target construct using a measurement model which is the most common one (i.e., the sum of the codes affected to the responses to the items). Indeed, when performing the systematic review on all MID methods, we found that almost all MID estimates were obtained using this measurement model [14].

Then, observed scores are scaled on a 0-100 continuum to express them on a common metric. Finally, using observed scores, MPCs are estimated for the aforementioned 70 retained methods on all the simulated datasets.

While the true populational MPC values are known in the same metric than the metric of the latent trait which is a continuum on the real line (and centered on 0 at the first time of measurement), MPC estimates are expressed on the score metric (i.e., a 0-100 continuum). Therefore, to estimate bias against true MPC values, we need to map these true values on the same metric than the metric of observed scores. True MPC values in the score metric are a function of the number of items J (3 different values), the number of response categories to the anchor K (3 different values) and the variance of the perceived change ($\sigma^2_{\theta^{pc(2)}}$). The variance of the perceived change is a function of β_1, β_2 and β_3 (26 different combinations) (see Eq 4). Thus, 234 different true MPC values are needed to be known in the metric of observed scores.

To get these true MPC values, we can simulate responses to the PRO items at two time of measurement of a sample of 100 000 individuals under the constraint of a true change equal to the simulated true MPC value in the metric of the latent trait ($\alpha^{(2)} = \frac{2\sigma_{\theta pc(2)}}{K-1}$). Then, the mean of observed scores are estimated for the dataset. The average difference of observed scores at each time of measurement is an estimate of the true MPC value in the score metric. This procedure is repeated for each of the 234 conditions to get all true MPC values in the score metric for the corresponding scenarios.

2. Analyses

As a first study, we will focus exclusively on the 378 scenarios where $\beta_1 = 1$ (the $\theta^{r(2)}$ is a perfect recovery of $\theta^{(1)}$) as this will allow analyzing the statistical performances of methods for MPC estimation under optimal conditions. These scenarios are the most appropriate ones to confirm or reject our working hypotheses.

First, our main objective is to analyze the overall statistical performances of the proposed methods for MPC estimation by pooling the results of the 378 aforementioned scenarios. Here, when using a method based on ROC analysis or predictive modeling to estimate the MPCs, we will first restrict analyses to the situation where the responses to the PGRC are used to classify patients as improved or not by using the "little change" group versus "the rest of the patients" (see II.1 and Figure 1).

For a given scenario, different indicators measuring the accuracy and variability of the MPC estimates for all the investigated methods will be used. These indicators will be averaged for the 500 replicates of the scenario. These indicators are: mean of the MPC estimates, variance of the MPC estimates, bias (difference between the estimate and the true MPC value), Mean Square Error (MSE, the square of the bias + the variance of the estimate), the Root Mean Square Error (RMSE). To answer our main objective, these different indicators will be averaged for the 378 investigated scenarios.

Then, we will study several secondary objectives. These analyses will be restricted on the subgroup of methods declared as unbiased after the previous analyses (unbiased is defined as an average bias for the 378 scenarios within a range of minus or plus 1 point around the true MPC value (in score metric)).

A first secondary objective is to investigate the influence of the 4 ways of using the responses to the PGRC for classifying patients as changed or not (see II.1) on the different indicators.

Then, another objective is to investigate the variability of the bias by the different conditions of the simulated scenarios (i.e., sample size, number of items…). To do so, for each globally unbiased MPC method, a linear model will be fitted with the estimated bias on the 378 scenarios as the dependent variable and the person and PRO characteristics (i.e., sample size, number of items, true change…) as the independent variables. Estimates of the model will be used to explore the robustness of the MPCs estimates under specific circumstances.

Finally, the same linear model will be performed for the least globally biased methods but using the MSE instead of bias as the dependent variable. This model will also be performed for unbiased methods. This analysis will help to investigate the possibility than, compared to unbiased methods, the loss of accuracy could in specific circumstances be compensated by an increase in precision (reflected by a smaller expected MSE value).

After analyses of the results under an ideal condition of a perfect recollection of the value of SC_{t1} at time 2, subsequent analyses will be focused on the influence of varying values of β_1, $\beta_2$and β_3 on the statistical performances of methods for MPC estimation. These analyses will be the focus of future works.

VI/ DISCUSSION

In this paper, we have thoroughly described a protocol for a large-scale simulation study investigating the statistical performances of many proposed estimators of the MIDs of PROs from the selection of methods to investigate, the conception of a theoretical model, the deduction of working hypotheses, the translation of the conceptual model into a simulation model to the statistical analysis plan. Several comments can be made.

First, the design of this simulation study presents several strengths. Indeed, to our knowledge, this is the first simulation study investigating in detail the statistical performances of many methods for MID estimations selected after a systematic review of the literature, including both anchor-based and distribution-based estimates. Second, our simulation model is the operationalization of an overarching conceptual model. This allows a clear distinction between a theory of structural relationships linking hypothesized agents which are supposed

to play a role in explaining the generation of data, and the operationalization of these relationships and agents into mathematical functions and variables. In addition, it allows to formally define what is the populational MPC value. This separation between the conceptual and simulation models helps to avoid a situation where the simulation process and analysis of the results lead to a tautology. Indeed, Monte-Carlo study results can be criticized if they obviously exhibit the expected results because of the operationalization of the simulation process and not because of the statistical performances of the investigated method(s). Or, at least, the data generation mechanism can favor some methods over others [23]. In addition, the distinction between a conceptual and simulation model allows an adequate examination of the assumptions underlying the generation of data, as well as its limitations [41]. Moreover, formally defining what is the populational MPC value has helped to deduce hypotheses on what can be appropriate MPC estimators; and has helped to propose an appropriate statistical analysis plan with clear primary and secondary objectives, avoiding a "data mining" strategy of analysis.

As aforementioned, three simulation studies have been performed to investigate the statistical performances of the estimator based on predictive modeling with sometimes a comparison against the estimator based on ROC analysis and Youden index [30–32]. As described in section III, it will be of great interest to see if the main results of these studies will be confirmed or not by the results of our simulation study. Nonetheless, the process of simulating data in these studies are quite different than the proposed process in our study. Thus, these differences need to be discussed as they can explain, in part, differences in results (if differences in results will be observed). Moreover, the simulation model used in Terluin and colleagues' studies has evolved from study to study (we will compare our simulation model with the one used in their last study). A first fundamental difference is the fact Terluin and colleagues' study simulate observed scored without simulating responses to items [32]. The simulation model used in this study assume these scores are obtained from items that are repeated parallel tests of the level of the construct of interest (i.e., all items measure the same construct, on the same scale, with the same item thresholds, and with the same amount of error) [32]. By simulating responses to items first, our model can be thought as more general, allowing us to investigate multiple issues that cannot be explored using the Terluin and colleagues' simulation model, such as the impact of choosing a specific measurement model, investigation of the impact of missing answers to the items, or investigation of the impact of lack of measurement invariance such as differential item functioning or response shift. A

second fundamental difference is the definition of the populational MID, and the operationalization of classifying patients as having changed or not. In the Terluin and colleagues' study, responses to a PGRC are not directly simulated [5]. Rather, patients are directly classified as having experienced a change or not by discretizing the distribution of thresholds reflecting a perfect relationship between change in the target construct and a hypothetical PGRC. Thus, in their study, structural relationships between the different agents involved in explaining responses to PRO items and to a PGRC are not operationalized, especially the relationships leading to the elicitation of perceived change values. Thus, their simulation model does not currently allow an explicit investigation of the influence of different agents (i.e., SCt1, SCt2, other contingencies) in explaining SCT_{1mem}, perceived change and ultimately responses to the PGRC. Last, a minor remark would be to notice that in Terluin and colleagues' study, baseline and final scores are simulated as uncorrelated [32].

As assumptions of our simulation model are explicit, there is room to discuss its simplifications and implications. First, the different agents involved in our conceptual model are operationalized in the simulation model as variables measured on continuous scales following normal distributions. A clear advantage of this assumption is the possibility to simulate individual values drawn from a distribution with a shape perfectly subsumed with only first and second moment parameters (i.e., mean and variance). If this assumption seems reasonable for the distribution of the target construct (normal distributions of scores (a measure of the target construct) are frequently observed on empirical data) and is the assumption usually done in psychometrics [42], it is a conjecture for the other agents operationalized in our simulation model. Then, as we could not formulate an *a priori* on the shape of the relationships between SC_{t1}, SC_{t2} and other contingencies (i.e., antecedents, catalyst, mechanisms) in explaining SCt_{1mem}, we were restricted to operationalize this phenomenological relationship as a quite simple model which is a first-degree polynomial function where each agent has an additive and independent contribution. Then, it must be noted our simulation model assumes symmetrical regularly spaced distributions (with equal variance) of PGRC thresholds (T_s) around the hypothetical point of no perceived change (see Eq 5 and Figure 6). An implication is a symmetrical value of the true MPC_+ and the MPC_-. Finally, our simulation model assumes that populational MPC values are not fixed values between individuals but are also a function of the number of categories to the PGRC (see Eq 5 and Eq 6). Specifically, when the number of response categories K increases, the MPC value decreases. The implication is we assume that, for a given individual, the MPC value is not an

intrinsic characteristic independent of the measurement process, but rather a value that is elicited because of the presentation of the PGRC. As such, our model assumes than when the granularity of response options to the PGRC increases, it elicits in people's mind a different MPC value than the one elicited when presenting a PGRC with less granularity. We propose this assumption is plausible, but it is a theoretical argument only.

Last, we can discuss the limits of our simulation study. First, our model does not allow to investigate the dependency of MPC values to baseline scores [32,43]. This phenomenon is frequently hypothesized based on empirical data (although a recent study from Terluin and al. shows this phenomenon can, in some cases, be wrongly hypothesized [32]). Second, our study does not investigate the occurrence of missing data, which is a frequent phenomenon in empirical studies about MID estimation [14,44]. Nonetheless, our simulation model could allow this investigation. Last, we assume the hypothesis of *longitudinal measurement invariance* of PRO scales holds. The violation of this hypothesis in the analysis of PRO longitudinal data is the subject of multiple empirical studies [6,45–47]. Because of a possible change in the meaning of a target construct over time, the hypothesis of longitudinal measurement invariance does not hold in some empirical situations. When this happens, observed change is not fully explained by target change: a phenomenon which is known as *response shift* [39]. The occurrence of a change in the meaning of the target construct between the two times of measurement could be an additional factor explaining bias in MPC estimations [48]. Our study assumes no response shift, therefore this impact of this phenomenon on bias cannot be currently assessed. Nonetheless, this could be addressed using our simulation model in further studies.

VII/ CONCLUSION

This study exposes a rigorous protocol for a large-scale simulation study, with a simulation model derived from a clear formal conceptual model, along with the assumptions, implications, and limits of the simulation process. Future analyses of the results of the study will help to confirm what are the appropriate methods for estimating the MPC of a PRO within ideal circumstances and explore the robustness of these methods under various conditions.

Because our work delineates its explicit assumptions, we also hope it will not be viewed as prescriptive or normative, but rather as a proposal of a rigorous framework for

fostering future research and generate debates on the definition of what is a responder definition threshold from the patient's perspective, what are the appropriate methods for estimating this threshold, and how future methods to estimate this threshold could be proposed.

ACKNOWLEDGHMENTS

This study is part of the MIDIPRES project, which was funded by the French National Agency for Research (ANR: "Agence Nationale de la Recherche", Jeunes Chercheurs 2016-2020 N° ANR-15-CE36-0003-01). The funders had no role in the design and conduct of the study; the collection, management, analysis, and interpretation of the data; the preparation, review, or approval of the manuscript; and the decision to submit the manuscript for publication.

REFERENCES

[1] P.M. Fayers, D. Machin, Quality of life: the assessment, analysis, and interpretation of patient-reported outcomes, 2nd ed, J. Wiley, Chichester ; Hoboken, NJ, 2007.

[2] D.L. Patrick, L.B. Burke, J.H. Powers, J.A. Scott, E.P. Rock, S. Dawisha, R. O'Neill, D.L. Kennedy, Patient-Reported Outcomes to Support Medical Product Labeling Claims: FDA Perspective, Value Health. 10 (2007) S125–S137. https://doi.org/10.1111/j.1524-4733.2007.00275.x.

[3] H.C.W. de Vet, ed., Measurement in medicine: a practical guide, Cambridge Univ. Press, Cambridge, 2011.

[4] J.C. Nunnally, I.H. Bernstein, Psychometric theory, 3rd ed, McGraw-Hill, New York, 1994.

[5] C.B. Terwee, J.D. Peipert, R. Chapman, J.-S. Lai, B. Terluin, D. Cella, P. Griffith, L.B. Mokkink, Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures, Qual. Life Res. (2021). https://doi.org/10.1007/s11136-021-02925-y.

[6] M.A.G. Sprangers, T.T. Sajobi, A. Vanier, N.E. Mayo, R. Sawatzky, L. Lix, F.J. Oort, V. Sébille, and the Response Shift - in Sync Working Group, Response shift in results of patient-reported outcome measures: A commentary to the Response Shift - in Sync Working Group Initiative, Qual. Life Res. Online ahead of print (2021).

[7] C.E. Schwartz, B.D. Rapkin, Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal, Health Qual. Life Outcomes. 2 (2004) 16.

[8] B.B. Reeve, K.W. Wyrwich, A.W. Wu, G. Velikova, C.B. Terwee, C.F. Snyder, C. Schwartz, D.A. Revicki, C.M. Moinpour, L.D. McLeod, J.C. Lyons, W.R. Lenderking, P.S. Hinds, R.D. Hays, J. Greenhalgh, R. Gershon, D. Feeny, P.M. Fayers, D. Cella, M. Brundage, S. Ahmed, N.K. Aaronson, Z. Butt, ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research, Qual. Life Res. (2013). https://doi.org/10.1007/s11136-012-0344-y.

[9] US Food and Drug Administration, Patient-Focused Drug Development Guidance Public Workshop. Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes, (2018).

[10] US Food and Drug Administration, Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, 2009.

[11] D.E. Beaton, C. Bombardier, J.N. Katz, J.G. Wright, A taxonomy for responsiveness, J. Clin. Epidemiol. (2001) 14.

[12] R. Jaeschke, J. Singer, G.H. Guyatt, Measurement of health status. Ascertaining the minimal clinically important difference, Control. Clin. Trials. 10 (1989) 407–415.

[13] A. Carrasco-Labra, T. Devji, A. Qasim, M.R. Phillips, Y. Wang, B.C. Johnston, N. Devasenapathy, D. Zeraatkar, M. Bhatt, X. Jin, R. Brignardello-Petersen, O. Urquhart, F. Foroutan, S. Schandelmaier, H. Pardo-Hernandez, Q. Hao, V. Wong, Z. Ye, L. Yao, R.W.M. Vernooij, H. Huang, L. Zeng, Y. Rizwan, R. Siemieniuk, L. Lytvyn, D.L. Patrick, S. Ebrahim, T.A. Furukawa, G. Nesrallah, H.J. Schünemann, M. Bhandari, L. Thabane, G.H. Guyatt, Minimal important difference estimates for patient-reported outcomes: A systematic survey, J. Clin. Epidemiol. 133 (2021) 61–71. https://doi.org/10.1016/j.jclinepi.2020.11.024.

[14] A. Vanier, P. Woaye-Hune, A. Toscano, V. Sébille, J.-B. Hardouin, What are all the proposed methods to estimate the Minimal Clinically Important Difference of a Patient-Reported Outcome Measure? A systematic review., in: Phila. 18-21 Oct 24th Annu. Conf. Int. Soc. Qual. Life, 2017.

[15] D. Revicki, R.D. Hays, D. Cella, J. Sloan, Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes, J. Clin. Epidemiol. 61 (2008) 102–109. https://doi.org/10.1016/j.jclinepi.2007.03.012.

[16] A.R. Sedaghat, Understanding the Minimal Clinically Important Difference (MCID) of Patient-Reported Outcome Measures, Otolaryngol. Neck Surg. 161 (2019) 551–560. https://doi.org/10.1177/0194599819852604.

[17] A.E. McGlothlin, R.J. Lewis, Minimal Clinically Important Difference: Defining What Really Matters to Patients, JAMA. 312 (2014) 1342–1343.

[18] J. Cohen, Statistical power analysis for the behavioral sciences, 2. ed., reprint, Psychology Press, New York, NY, 2009.

[19] G.R. Norman, J.A. Sloan, K.W. Wyrwich, The truly remarkable universality of half a standard deviation: confirmation through another look, Expert Rev. Pharmacoecon. Outcomes Res. 4 (2004) 581–585. https://doi.org/10.1586/14737167.4.5.581.

[20] K.W. Wyrwich, W.M. Tierney, F.D. Wolinsky, Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life, J. Clin. Epidemiol. 52 (1999) 861–873.

[21] K.W. Wyrwich, W.M. Tierney, F.D. Wolinsky, Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire, Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil. 11 (2002) 1–7.

[22] N.K. Leidy, K.W. Wyrwich, Bridging the Gap: Using Triangulation Methodology to Estimate Minimal Clinically Important Differences (MCIDs), COPD J. Chronic Obstr. Pulm. Dis. 2 (2005) 157–165. https://doi.org/10.1081/COPD-200050508.

[23] T.P. Morris, I.R. White, M.J. Crowther, Using simulation studies to evaluate statistical methods, Stat. Med. 38 (2019) 2074–2102. https://doi.org/10.1002/sim.8086.

[24] A.M. Hinds, T.T. Sajobi, V. Sebille, R. Sawatzky, L.M. Lix, A systematic review of the quality of reporting of simulation studies about methods for the analysis of complex longitudinal patient-reported outcomes data, Qual. Life Res. 27 (2018) 2507–2516. https://doi.org/10.1007/s11136-018-1861-0.

[25] M. Blanchin, A. Guilleux, J.-B. Hardouin, V. Sébille, Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level: A simulation study, Stat. Methods Med. Res. 29 (2020) 1015–1029. https://doi.org/10.1177/0962280219884574.

[26] A. Rouquette, J.-B. Hardouin, J. Coste, Differential Item Functioning (DIF) and Subsequent Bias in Group Comparisons using a Composite Measurement Scale: A Simulation Study, J. Appl. Meas. 17 (2016) 312–334.

[27] É. de Bock, J.-B. Hardouin, M. Blanchin, T. Le Neel, G. Kubis, V. Sébille, Assessment of score- and Rasch-based methods for group comparison of longitudinal patient-

reported outcomes with intermittent missing data (informative and non-informative), Qual. Life Res. 24 (2015) 19–29. https://doi.org/10.1007/s11136-014-0648-1.

[28] A.M. Hinds, T.T. Sajobi, V. Sebille, R. Sawatzky, L.M. Lix, A systematic review of the quality of reporting of simulation studies about methods for the analysis of complex longitudinal patient-reported outcomes data, Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil. 27 (2018) 2507–2516. https://doi.org/10.1007/s11136-018-1861-0.

[29] A. Vanier, V. Sébille, M. Blanchin, J.-B. Hardouin, The minimal perceived change: a formal model of the responder definition according to the patient's meaning of change for patient-reported outcome data analysis and interpretation, BMC Med. Res. Methodol. 21 (2021) 128. https://doi.org/10.1186/s12874-021-01307-9.

[30] B. Terluin, I. Eekhout, C.B. Terwee, H.C.W. de Vet, Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis, J. Clin. Epidemiol. 68 (2015) 1388–1396. https://doi.org/10.1016/j.jclinepi.2015.03.015.

[31] B. Terluin, I. Eekhout, C.B. Terwee, 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, J. Clin. Epidemiol. 83 (2017) 90–100. https://doi.org/10.1016/j.jclinepi.2016.12.015.

[32] B. Terluin, E.M. Roos, C.B. Terwee, J.B. Thorlund, L.H. Ingelsrud, Assessing baseline dependency of anchor-based minimal important change (MIC): don't stratify on the baseline score!, Qual. Life Res. (2021). https://doi.org/10.1007/s11136-021-02886-2.

[33] D.A. Redelmeier, G.H. Guyatt, R.S. Goldstein, Assessing the minimal important difference in symptoms: A comparison of two techniques, J. Clin. Epidemiol. 49 (1996) 1215–1219. https://doi.org/10.1016/S0895-4356(96)00206-5.

[34] F. Tubach, G.A. Wells, P. Ravaud, M. Dougados, Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues, J. Rheumatol. 32 (2005) 2025–2029.

[35] D. Turner, H.J. Schünemann, L.E. Griffith, D.E. Beaton, A.M. Griffiths, J.N. Critch, G.H. Guyatt, Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference, J. Clin. Epidemiol. 62 (2009) 374–379. https://doi.org/10.1016/j.jclinepi.2008.07.009.

[36] B.D. Rapkin, C.E. Schwartz, Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift, Health Qual. Life Outcomes. 2 (2004) 14. https://doi.org/10.1186/1477-7525-2-14.

[37] R. Tourangeau, L.J. Rips, K.A. Rasinski, The psychology of survey response, Cambridge University Press, Cambridge, U.K.; New York, 2000.

[38] L. Engel, D.E. Beaton, Z. Touma, Minimal Clinically Important Difference. A review of Outcome Measure Score Interpretation., Rheum. Dis. Clin. N. Am. 44 (2018) 177–188. https://doi.org/10.1016/j.rdc.2018.01.011.

[39] A. Vanier, F.J. Oort, L. McClimans, N. Ow, B.G. Gulek, J.R. Böhnke, M. Sprangers, V. Sébille, N. Mayo, and the Response Shift - in Sync Working Group, Response shift in patient-reported outcomes: definition, theory, and a revised model, Qual. Life Res. (2021). https://doi.org/10.1007/s11136-021-02846-w.

[40] K.B. Christensen, S. Kreiner, M. Mesbah, eds., Rasch Models in Health: Christensen/Rasch Models in Health, John Wiley & Sons, Inc., Hoboken, NJ USA, 2012. https://doi.org/10.1002/9781118574454.

[41] J.B. Grace, D.R. Schoolmaster, G.R. Guntenspergen, A.M. Little, B.R. Mitchell, K.M. Miller, E.W. Schweiger, Guidelines for a graph-theoretic implementation of structural equation modeling, Ecosphere. 3 (2012) art73. https://doi.org/10.1890/ES12-00048.1.

[42] K.A. Bollen, R. Hoyle, Latent Variables in Structural Equation Modeling, in: Handb. Struct. Equ. Model., Guilford Press, New-York, 2012: pp. 56–67.

[43] C.B. Terwee, L.D. Roorda, J. Dekker, S.M. Bierma-Zeinstra, G. Peat, K.P. Jordan, P. Croft, H.C.W. de Vet, Mind the MIC: large variation among populations and methods, J. Clin. Epidemiol. 63 (2010) 524–534. https://doi.org/10.1016/j.jclinepi.2009.08.010.

[44] P. Woaye-Hune, J.-B. Hardouin, P.A. Lehur, G. Meurette, A. Vanier, Practical issues encountered while determining Minimal Clinically Important Difference in Patient-Reported Outcomes, Health Qual. Life Outcomes. 18 (2020) 156.

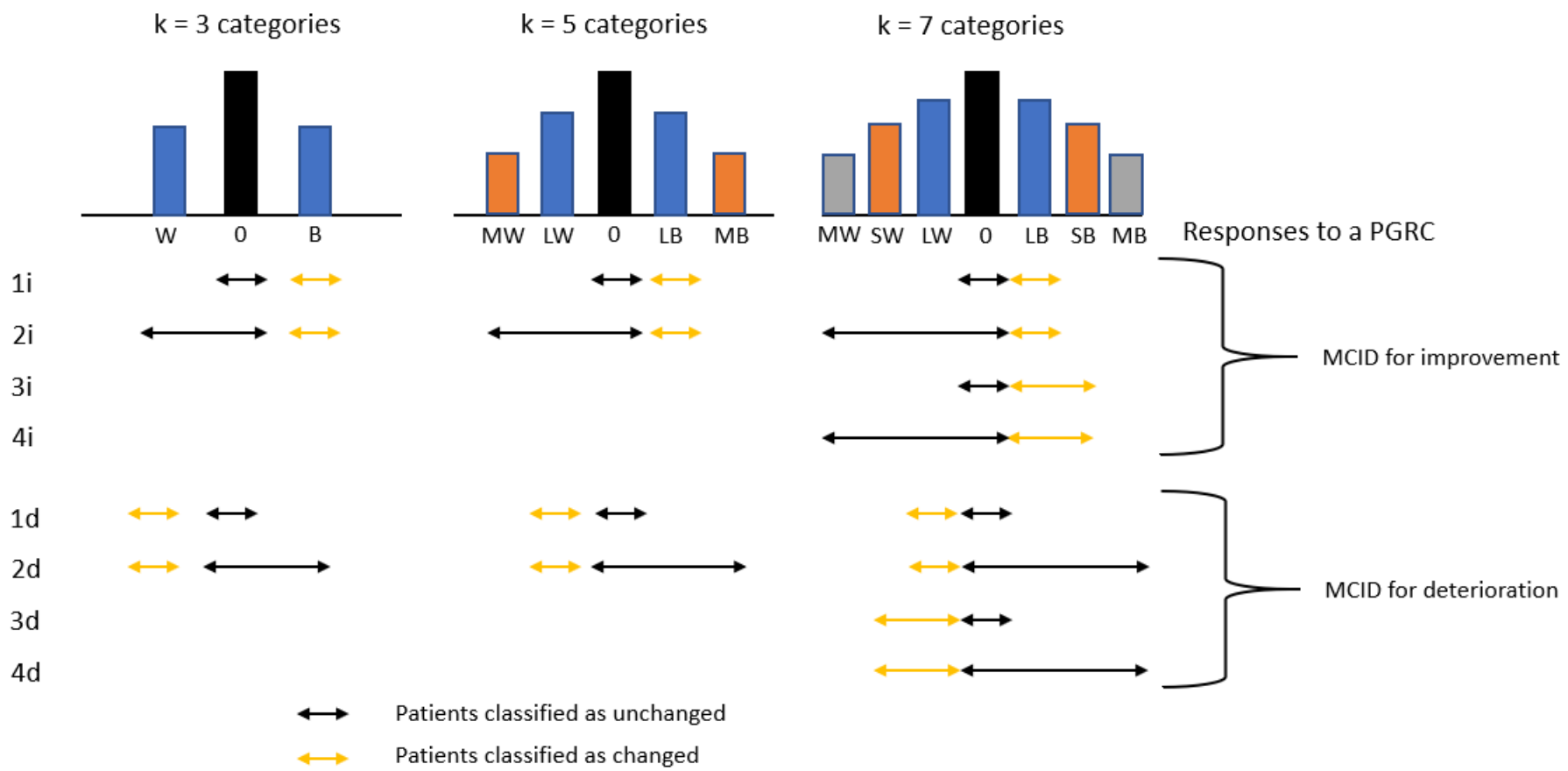
[45] F.J. Oort, M.R.M. Visser, M.A.G. Sprangers, An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery, Qual. Life Res. 14 (2005) 599–609.

[46] M. Salmon, M. Blanchin, C. Rotonda, F. Guillemin, V. Sébille, Identifying patterns of adaptation in breast cancer patients with cancer-related fatigue using response shift analyses at subgroup level, Cancer Med. 6 (2017) 2562–2575. https://doi.org/10.1002/cam4.1219.

[47] C.E. Schwartz, R. Bode, N. Repucci, J. Becker, M.A.G. Sprangers, P.M. Fayers, The clinical significance of adaptation to changing health: a meta-analysis of response shift, Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil. 15 (2006) 1533–1550. https://doi.org/10.1007/s11136-006-0025-9.

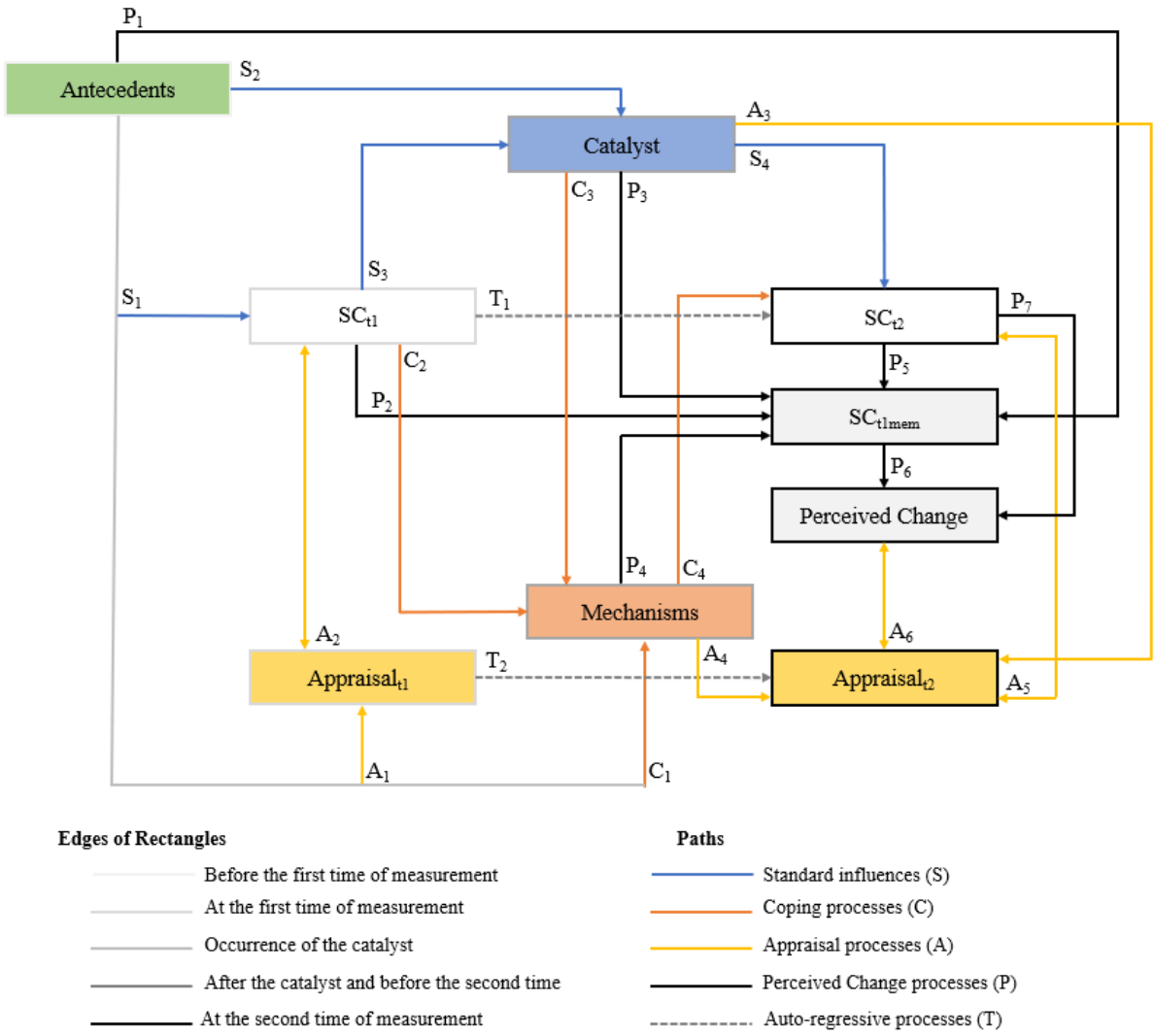
[48] C.E. Schwartz, V.E. Powell, B.D. Rapkin, When global rating of change contradicts observed change: Examining appraisal processes underlying paradoxical responses over time, Qual. Life Res. 26 (2017) 847–857. https://doi.org/10.1007/s11136-016-1414-3.

Figure 1. Four different ways of classifying patients as changed or unchanged using the responses to a PGRC for estimating a MID using ROC analysis or predictive modeling



Notes: 0 = Stable, W = Worse, B = Better, M = Much, L = Little, S = Somewhat

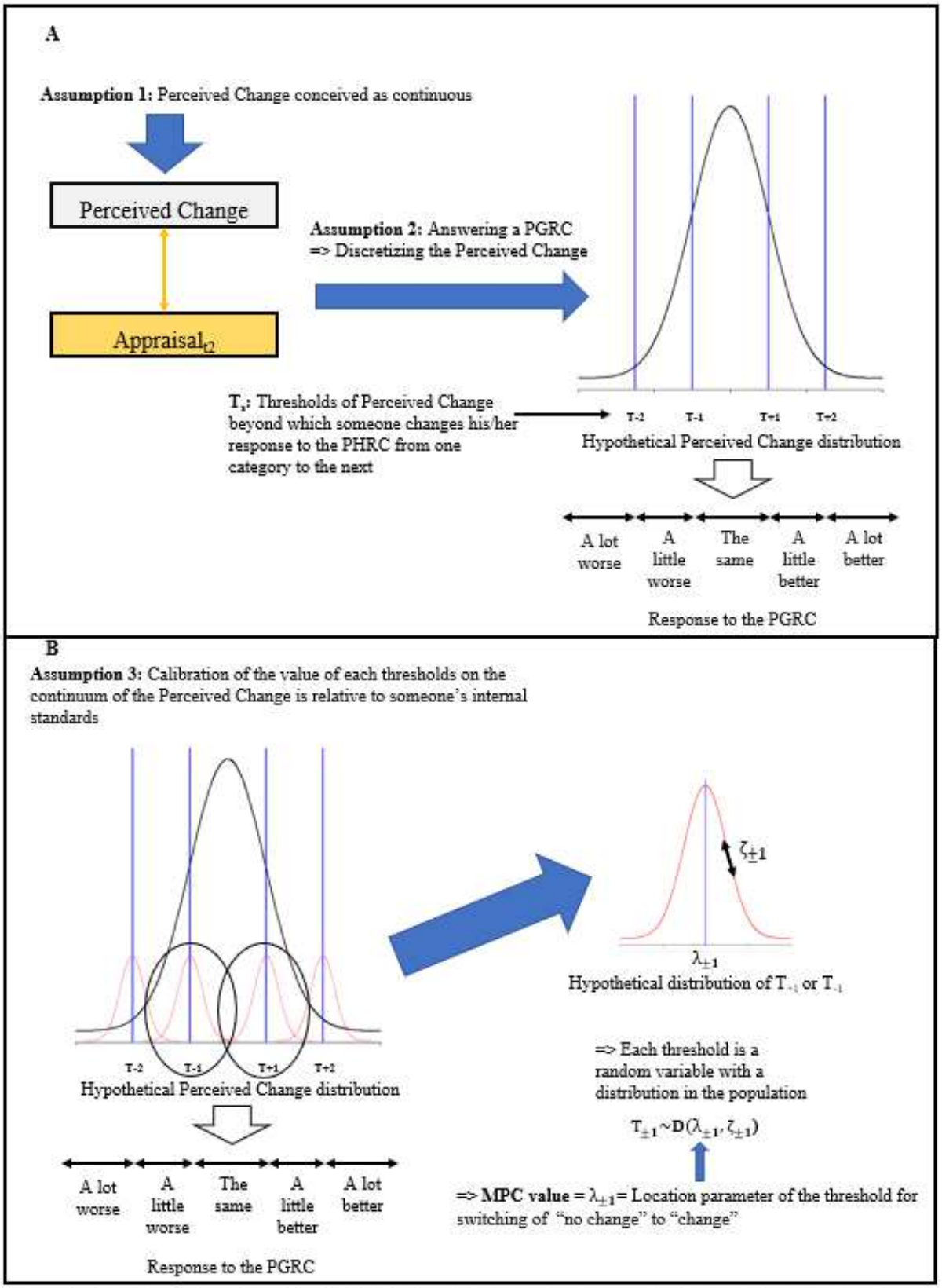
Figure 2. A theoretical model depicting the agents engaged when someone rates his/her level on a given PRO at two times of measurement and answers a PGRC at the second time (Source: Vanier at al. (2021) BMC Medical Research Methodology)



Abbreviations: SC: Subjective Construct, t1: first time, t2: second time, SC_{t1mem}: Remembered baseline level of the subjective construct

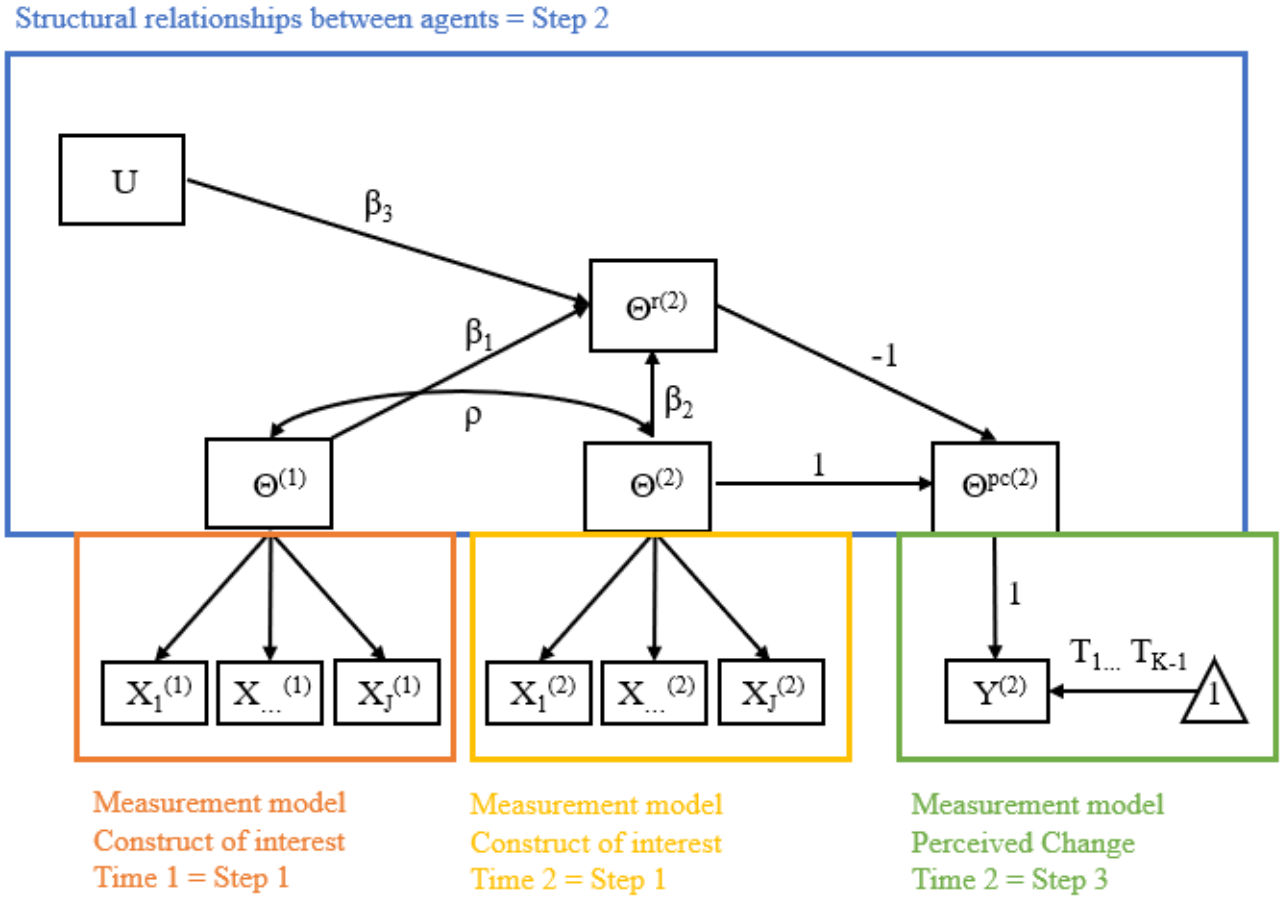
Notes: Unidirectional arrows are cause-effect relationships. Bidirectionnal arrows are correlations.

Figure 3. A definition of the MPC as a statistical parameter in the population. A: The measurement of perceived change by a PGRC. B: Defining the value of the MPC (Source: Vanier at al. (2021) BMC Medical Research Methodology)



Abbreviations. PGRC: Patient Global Rating of Change, MPC: Minimal Perceived Change. **Notes**. This representation depicts a PGRC with 5 response categories (so 4 thresholds) as an example. For simplicity, hypothetical populational distributions are represented using gaussian curves but the real shape of these distributions could be another one.

Figure 4. A simulation model for generating responses to items of a given PRO at two times of measurement and a response to a PGRC at the second time of measurement



Notes: $X_j^{(t)}$: Item j (j=1;…;J) at time t (t=1;2). $\theta^{(t)}$: Construct of interest at time t. ρ: Correlation between $\theta^{(1)}$and $\theta^{(2)}$. $\theta^{r(2)}$: Remembered latent trait at baseline, this construct is elicited at the second measurement occasion. U: a random variable with a known populational distribution representing the influence of other contingencies (e.g., influence of the catalyst) on the process of remembering/inferring $\theta^{r(2)}$. β_1, β_2, β_3: The relative importance of each of the three components ($\theta^{(1)}$, $\theta^{(2)}$, U) in determining $\theta^{r(2)}$. By constraint, $\beta_1+\beta_2+\beta_3=1$. $\theta^{pc(2)}$: Perceived Change. It is the difference between $\theta^{(2)}$and $\theta^{r(2)}$. $Y^{(2)}$: The response to the PGRC. T_k: For an anchor with K (k=1,…,K) categories, the k^{th} threshold for discretizing the Perceived Change into a categorical response. For an PGRC with K categories, there exists K-1 thresholds. Each threshold has a known populational distribution with mean and variance. The means of the $(K+1)/2^{th}$ threshold and the $(K-1)/2^{th}$ threshold are the populational known MPC values for improvement and deterioration, respectively.

Figure 5. A hypothetical well-adapted questionnaire with 5 dichotomous items. Each vertical bar is an item threshold. The global item distribution covers the latent trait distribution

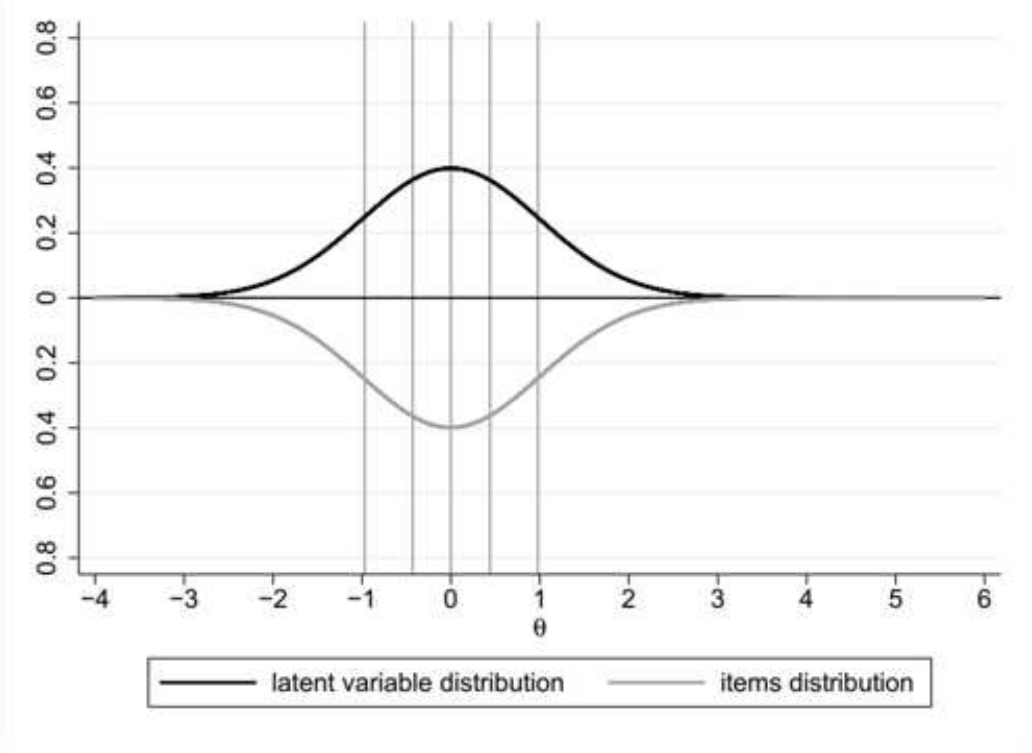
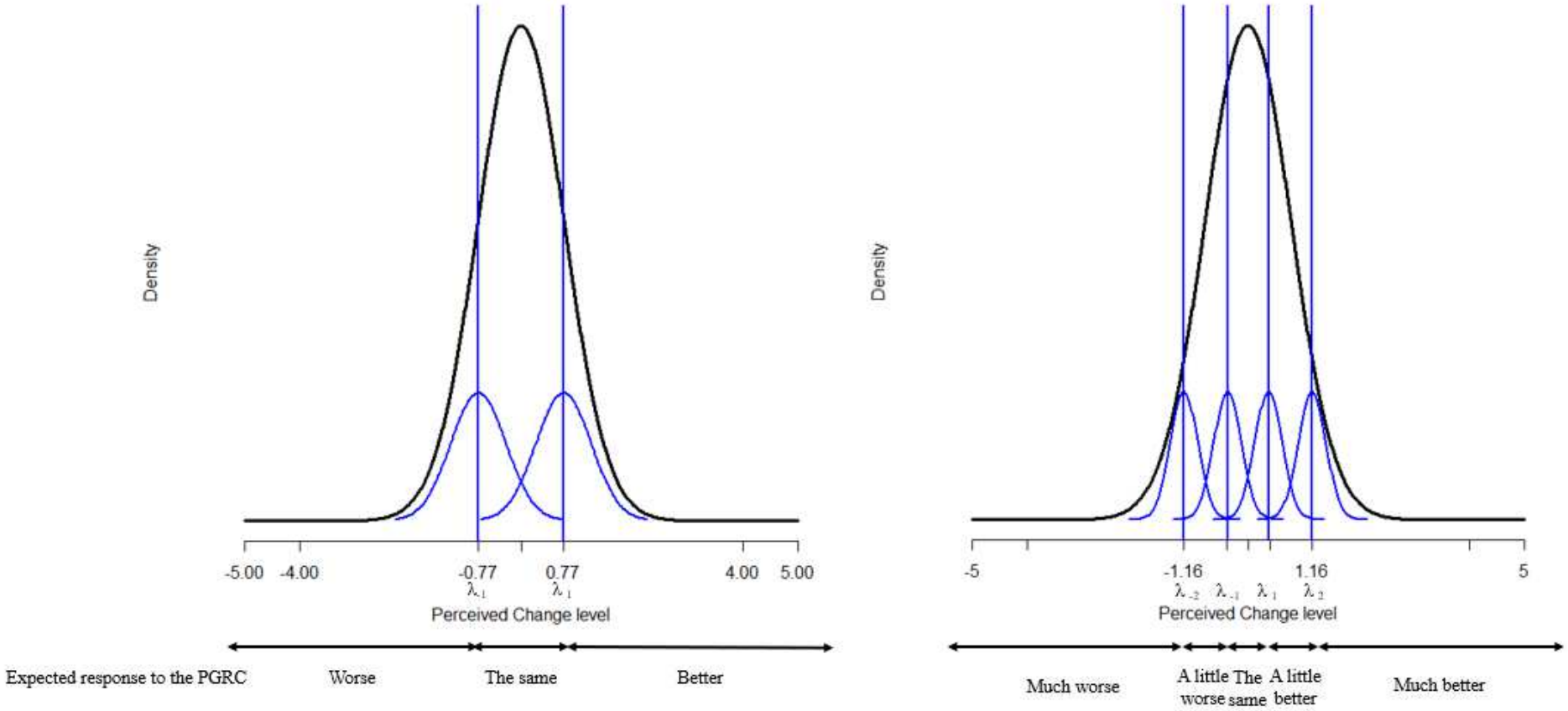


Figure 6. An example of the discretization of the Perceived Change into a response to the PGRC according to known true populational MPC value (for k=3 and k=5, when the SC_{t1} is perfectly recollected, with an absence of true change and $\zeta_s^2 = \left(\frac{1}{K-1}\right)^2$)



Notes: Black curves are the distribution of the Perceived Change. Blue curves are the distribution of the thresholds. Blue lines are mean thresholds values.

Table 1. Methods for estimating MIDs retained in the simulation study (specificities can be found in section II.1, Figure 1 and Supplementary eText 1)

Main category	Subtype of estimator		Specific characteristics to certain methods	
Anchor-based methods*	Mean of the "little change" group Mean of the "little change" group minus Mean of the "stable" group 75th percentile of the "little change" group			
	Best threshold using ROC analysis	**K** 1. 3 2. 5 3. 7	**4 ways for classifying patients into two groups** 1. "Little change" versus "stable" only 2. "Little change" versus "rest of the patients" 3. "Little + somewhat change" versus "stable" only 4. "Little + somewhat change" versus "rest of the patients"	**5 criteria** 1. Euclidian 2. Youden 3. Sp = 80% 4. Sp = 100% 5. Se = 100%
	Best threshold using predictive modeling			**2 estimators** 1. Crude 2. Adjusted
Distribution-based methods	Based on Cohen's Effect Size			**3 estimators** 1. 0.2 effect size 2. 0.5 effect size 3. 0.8 effect size
	Based on Standardized Response Mean (SRM)			**3 estimators** 1. 0.2 SRM 2. 0.5 SRM 3. 0.8 SRM
	Based on measurement error			**3 estimators** 1. 1 Standard Error of Measurement 2. 1.96 Standard Error of Measurement 3. 1 Minimal Detectable Change
	Based on range of scale of change in scores			**2 estimators** 8% of the ranger of the change 7% of the total change possible

Notes: K= Number of response categories to the PGRC, Sp = Specificity, Se = Sensitivity

***** All anchor-based methods can be used for estimating an MID for improvement and a MID for deterioration

Table 2. Varying characteristics in the different scenarios explored by the simulation study

Characteristics	Values
Person characteristics	
Sample size (N)	200, 500, 1000
True change in the target construct ($\alpha^{(2)}$)	-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8
The importance of $\theta^{(1)}$ on explaining $\theta^{r(2)}$ (β_1)	0, 0.1, 0.2, 0.3, 0.5, 1
The importance $\theta^{(2)}$ of on explaining $\theta^{r(2)}$ (β_2)	0, 0.1, 0.2, 0.3, 0.5
The variance of the PGRC thresholds ζ_s^2	$\left(\frac{1}{K-1}\right)^2, \left(\frac{0.5}{K-1}\right)^2$
PRO characteristics	
Number of response categories to the PGRC (K)	3, 5, 7
Number of items of the PRO(J)	5, 10, 20