Introduction to the special section

Trigg, Andrew; Lenderking, William R; Boehnke, Jan R

DOI:
10.31234/osf.io/75a8c

Publication date:
2023

Editorial

Introduction to the special section

"Methodologies and Considerations for Meaningful Change"

Andrew Trigg[1], William R. Lenderking[2], Jan R. Boehnke[3]

[1] Medical Affairs Statistics, Bayer plc, Reading, United Kingdom

[2] Patient-Centered Research, Evidera, Bethesda, MD, USA

[3] School of Health Sciences, University of Dundee, Dundee, UK

Conflicts of interest

AT declares current employment by Bayer plc and prior employment by Adelphi Values and DRG. AT was purposefully not involved in editorial decisions regarding papers submitted by any employees of these companies, and declares no other financial or non-financial conflicts of interest.

WRL declares current employment by Evidera, formerly employed at Pfizer (2000-2007). WRL was not involved in editorial decisions involving papers submitted by current Evidera or Pfizer colleagues, and declares no other financial or non-financial conflicts of interest.

JRB is Co-Editor in Chief of Quality of Life Research and declares no financial or non-financial conflicts of interest.

This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is published in Quality of Life Research, and is available online at https://doi.org/10.1007/s11136-023-03413-1

Corresponding author:

Jan R. Boehnke

University of Dundee

School of Health Sciences

11 Airlie Place

Dundee DD1 4HJ

United Kingdom

The determination of what constitutes a 'meaningful change' on a health outcome measure remains controversial in both methodological and applied research. Motivated by the question of how to understand the efficacy and effectiveness of interventions or the natural history of conditions better (1,2), the concept builds on the widely held belief that statistical significance in itself is not sufficient to establish a treatment benefit (3,4). Since health-related quality of life (HRQL) research should reflect patients' perceptions and evaluations, the topic is of immense theoretical, statistical, and practical relevance. It was therefore timely to offer a space to present discussions, methods, and questions related to this topic, even as new methods and interpretive standards emerge.

In collaboration with the Psychometrics Special Interest Group of the International Society for Quality of Life Research (ISOQOL) the editor-in-chief (JRB) developed a call for papers and selected the editorial team (AT and WRL) to take this special issue forward. Submissions closed in April 2021 and invited submissions exploring existing and novel methods for defining meaningful change thresholds for clinical outcome assessments such as patient- or clinician-reported outcome measures. A simulation dataset, described below, was also provided to encourage researchers to evaluate different methods using the same data. The main aim of the special section was to collate a series of methodological and applied articles reflecting current thinking and developments in meaningful change research. And we also wanted to encourage the practice of explicitly stating whether thresholds are intended to support between-group, within-group or within-individual interpretations (3,5–7).

For this special section, we broadly define "meaningful change research" as the determination of guidelines for interpretation of the perceived meaning of health outcome score changes or differences based on the patients' (or: the target population's) perception. For a particular score difference (often described as a "threshold") to indicate a "meaningful change" over time, (i) patients (or an appropriate proxy) need to have described this score difference as directional (e.g., improved or deteriorated); and (ii) to a degree that reflects in their eyes a meaningful difference from the previous state (see for example) (3,4). A variety of methods are used to operationalize this, including anchor-based methods or qualitative evaluations of score differences that are perceived as meaningful (8).

When working towards concrete operationalizations, the level, type, and magnitude of change need to be specified. For example, it is likely inadmissible to use change thresholds based on group differences to interpret differences between individuals or within individuals over time (7), although this may be a common practice. Table 1 provides an overview of these three key considerations when classifying change and we point out three examples:

- Minimal within-individual change over time: the smallest amount of change over time a given person must show on an individual level in order to be regarded as having a meaningful change (1B, 2B, 3A);
- Minimal between-group difference in change over time: the smallest difference between the changes of one group versus another group that are considered meaningful (1A, 2C, 3A);
- Minimal within-group change over time: the smallest amount of change over time a group of people must show in order to be regarded as having had a meaningful change (1A, 2B, 3A).

Other combinations such as cross-sectional between-individual differences are also made in practice (3), in addition to 'larger than minimal' thresholds (9). Similarly, while some definitions focus on changes that ‘warrant a change in a patient’s management’ (12), we do not consider this to be a necessity, as some studies (natural history) do not involve treatment evaluations, yet still must establish a meaningful change. Finally, we consciously avoid the use of specific terms such as ‘minimal clinically important difference’ (13) or ‘minimally important change’ within this editorial (4), given these terms have been used interchangeably to describe a range of the combinations arising from Table 1. Standardized terminology is more likely to be achieved through a consensus-based approach in a large group such as the SISAQOL-IMI (14). Until consensus is achieved, it is essential for clarity of communication that all dimensions in Table 1 are clarified in the description of a threshold, e.g., “minimal within-individual change over time”.

Insert Table 1 about here

The special section is split into two parts: the first focuses on meaningful change using clinical anchors, the second one presents papers based on what are often called "distribution-based" approaches. Distribution-based approaches are typically described as (i) using measures of cross-sectional or longitudinal (often inter-individual) variability in order to define (ii) a minimal score difference that would be seen as exceeding the level of measurement error (or otherwise nuisance or negligible variability) given a particular psychometric model (15). These thresholds have no connection to (external) evaluations of "meaningfulness" of that particular score difference. It is for this reason that regulators such as FDA have historically stated that distribution-based approaches cannot be used as the sole basis for establishing a responder definition (16). Instead, the assumption is that score differences that are greater than measurement error are due to a more systematic factor or factors, hence the inference of meaning. Their singular advantage in this context is that they do not depend on finding a suitable external clinical anchor, which can be challenging for some applications, but can be calculated solely using data from the measure being evaluated. In contrast, an index of meaningful change would offer information about 'meaningfulness' by either providing information about the connection to a criterion of change or by offering a clear content-based operationalization of meaningfulness (be it qualitative or quantitative). However, when such a criterion is not available, distribution-based methods can be useful. Furthermore, in this special section, the submissions were of high quality, and their inclusion offers the opportunity to contrast the approaches, and the contribution of these methods is too important to leave out of a special section such as this. Additionally, they have an established history of use for the study of individuals over time (i.e., idiographic research) to complement trends at the group level (i.e., nomothetic research) (5,17,18).

Finally, we want to thank Pip Griffiths (Digital Medicine Society; IQVIA; SeeingTheta) for providing the simulated dataset that two articles used to illustrate their approaches (19,20), and which could be interesting for readers to explore some of the issues raised in this special section further. The simulated dataset comprises responses to the twelve-item ‘Simulated Disease Questionnaire’ for 2,000 individuals at four time points. The items have four response categories where higher scores indicate worse health (graded response model). Responses to a seven-category transition rating (i.e., global impression of

change) were also simulated at the follow-up time points (for more details please refer to https://osf.io/khmzg/).

THE SPECIAL SECTION

The response to the call was enthusiastic, with twenty-seven submissions exploring a range of conceptual and practical issues, of which fifteen are now brought together in this special section. Ten of these papers focus on *meaningful change*, and five papers and two letters address *distribution-based indices*. The focus of each paper, in terms of meaningful change versus distribution-based indices, and further classification on the level and type of threshold, is provided in Table 2. Two things are clear from this table. First, most papers focus on within-individual change over time. Second, several papers on meaningful change did not precisely specify the magnitude of change (minimal versus greater). For one of these cases, meaningful change was instead conceptualized in terms of hypothetical patient-perceived treatment success (21). For another paper (22) specifying the magnitude, authors used the terms minimal to reflect ratings of 'a little better' and meaningful to reflect ratings of 'better' and 'much better'. We recommend future papers are clearer in terms of the intended magnitude, but note that the two options for the magnitude dimension in Table 1 are not exhaustive where options such as patient-perceived treatment success can be of interest.

Insert Table 2 about here

Setting the scene for the first part of the special section is a report of an online survey regarding how clinicians from different disciplines determine individual-level meaningful change on patient reported outcome measures (PROMs) (23). The authors investigated how oncology or mental health clinical care

providers who used PROMs in the USA determine whether a patient's symptoms have changed. Most commonly, clinicians compared two consecutive scores, without a visual aid; the use of normative scores was uncommon. This research highlights the importance of aligning meaningful change research with current practice, but also that education in the value of interpretative tools is warranted.

The papers in this section investigate the use of anchors for the derivation of meaningful change thresholds. Anchor-based methods are the most widely applied method for estimating meaningful change, but this does not mean they are without problems. In the second paper of this section (24), the authors highlight and discuss five important issues with anchors that should be kept in mind, rather than viewing anchor-based approaches as a perfect gold standard. This article serves as a helpful collection of methodological issues to consider when reading the collected papers. The third paper illustrates a fundamental practical question when determining meaningful change thresholds, but likely also for any threshold determination (10): how scoring rules and ranges limit the usability of group-level minimal important differences in individual-level responder definitions. Based on the example of the EORTC QLQ-C30 subscales, the authors illustrate how the commonly used 10-point change may be misleading, as due to scaling, an individual cannot actually be measured with a 10-point change on any scale. They present considerations (their Figure 2) to further support responder threshold selection.

Moving to investigations of the effectiveness of study design and analysis approaches, the fourth paper (25) reports the results of a simulation study to evaluate the importance of the strength of the correlation between the anchor and the clinical outcome assessment measuring change, varying the impact of sample size, change score variability, and anchor correlation strength on the estimation of the meaningful change threshold at the individual and group level. Using receiver operator characteristics and logistic regression analyses, they show that sample size and change score variability are key factors impacting the required anchor correlation, but using an 'acceptable' cut-off of > 0.30 was often insufficient for accurate estimates of individual meaningful change thresholds, and always insufficient for group changes. The fifth paper (19) builds on the simulation dataset that accompanied this call to address the problem that traditional methods of evaluating within-individual change ignore the effects of floor/ceiling effects and measurement error in PROM scores and global (transition) ratings. The team

combined the use of a longitudinal graded response model with a transition item to measure latent change. The method produced tighter estimates of meaningful change when compared to traditional methods, with the methods overlapping most when the proportion of responders was about 50% of participants. Extensions of this approach show promise for a range of applications (26,27). The final simulation study in the first part (28) casts a view forward to the papers on distribution-based thresholds, as the team evaluated the effects of sample characteristics commonly observed in clinical trials on four anchor-based threshold selection procedures and two distribution-based ones. In a large simulation design, they found that both methodological choices and clinical characteristics exert influence on the results and conclusions, and they suggest prioritising study designs with strongly responsive endpoints in settings with about 50% anchor-based responders.

Moving to empirical papers exploring questions of meaningful change, one team explored if, how and when meaningful change in depressive symptoms occurred during a period of four months through three data sources (18): weekly questionnaires, qualitative reports, and ecological momentary assessment (EMA; five prompts per day). The 'if' was assessed in terms of measurement error (weekly level), perceived meaningfulness (qualitative), and statistically significant changes in the modelled trajectory of symptoms. The distinction between sudden and gradual change (the how) and when this occurred varied considerably between methods. This research will help others evaluate what information each method can provide, alone or in combination, when designing studies to assess health changes. It also points to the potential of EMA and experience sampling to increase patient-centeredness and granularity when collecting HRQL data (29). The use of multiple data sources also plays a key role in the three papers concluding this section. One team (22) sought to evaluate the validity of a rheumatoid arthritis flare questionnaire by examining minimal and meaningful within-individual change using three anchors: patient global ratings, physician global ratings, and using a disease activity index in patients with rheumatoid arthritis. They found that patients were most likely to report meaningful improvement, physicians were most likely to report meaningful worsening, with changes in either direction on the disease activity index least likely to be classified as meaningful. Another team (21) utilized a clinicians-then-patients qualitative interview methodology to understand patient priorities for treatment and a

threshold to declare treatment success for adult and adolescent patients with alopecia areata and ≥ 50% scalp hair loss. This paper details the novel qualitative method of explicitly incorporating patient input into the definition of an individual change threshold and the endpoint of %hair loss. The authors documented that due to extensive discussions online by patients about hair loss issues, they were able to make appropriate ratings of their hair loss that were largely consistent with values provided by clinical experts. The first part ends with a qualitative study to define meaningful change in physical function after weight-loss (30). The team conducted a qualitative study to evaluate how much weight loss would be meaningful hypothetically for overweight and obese individuals, if they were to lose weight. These individuals all agreed that a $\geq$ 10% weight loss would be associated with a meaningful improvement in their physical functioning, and that a one-point change at the item level of two HRQL instruments would represent a noticeable change.

The papers in the second part of the special section focus on *distribution-based indices*. The papers explore how these indices and precision of their recovery are affected by different definitions of the error variance, distributions, and level of uncertainty. The first paper (31) builds upon previous work by the authors (5) proposing approaches for the identification of treatment responders, providing further justification and elaboration for the use of the coefficient of repeatability (also known as the 'smallest real difference' (32) or 'minimally detectable change' (15)) for within-individual interpretations of statistically significant change. However, rather than focusing on the conventional $p < 0.05$ threshold, the authors explore more liberal thresholds. This article serves as a helpful reminder that significance levels are not fixed, where less strict (i.e., smaller) thresholds will be sufficient in some scenarios. In addition, the paper has two letters attached to it in this same issue, which discuss the interpretation of the attached statistical significance level and the applicability of the index to individual change classification, which are also of interest for other indices and their interpretation. The second paper (33), focuses also on a version of the reliable change index and compares its use based on classical test theory and item response theory. Classical test theory assumes measurement error is constant across the scale range, but item response theory relaxes this assumption. The authors compare these approaches to detect change beyond measurement error, where the item response theory-based thresholds fluctuate above or

below the fixed classical test theory threshold in accordance with baseline score. Their Table 4 presents an overview of thresholds for PROMIS shortform users within oncology. Using item response models, another team (34) proposes a method for increasing the precision of measurement of within-individual change. They build on existing approaches to quantify the error associated with individual scores derived from item response theory analyses: using plausible values, the precision of scores across the spectrum of theta (severity of underlying trait) can be incorporated. This can increase the accuracy of measuring intra-individual changes, which is very useful in individuals (for example) with chronic illness who need to be monitored repeatedly over time and provides an extension to more typical distribution-based methods.

All PROM scores are subject to measurement error and using raw individual change scores does not account for this fact. The last two papers in the special section use regression and predictive frameworks to derive change metrics that also allow to quantify the uncertainty associated with the estimate. One team (35) presents alternatives to the raw change scores that were developed over 50 years ago (36,37), but have so far not been widely used or explored within patient-reported outcome research. The two approaches provide estimates of an individual's true gain after incorporating measurement error, which have both conceptual advantages and greater sensitivity compared to raw change scores. The final paper of the special section (20), compares three distribution-based methods: the reliable change index, one of its variants, and Bayesian regression models that regress post-scores on pre-scores to identify group-level change over time. The article shows that there are only small differences between the methods in detecting change when PROM reliability is high, but none of them outperforms all others if that is not the case. The article offers a technical discussion that compares advantages and disadvantages of these approaches.

EDITORIAL COMMENTARY

In closing, we want to take the opportunity to highlight three topics that struck us when reading and editing the papers. A first observation is that anchor-based methods for within-individual guidelines

should be based on finding a threshold separating 'no change' and 'changed' groups on the anchor. The notion of locating a threshold, lying along a continuum of perceived change, is supported by recent research (38). As individuals will vary in their personal threshold, many methods use the mean of these individual threshold locations or derive otherwise a threshold aggregated across individuals (e.g., receiver operating characteristic curves, logistic regression, discriminant analysis; (4,39)). Similarly, the longitudinal item response model presented within this special section is designed to estimate the location of this threshold (19). Therefore, from a theoretical standpoint we view anchor-based methods such as receiver operating characteristic curves, logistic regression, discriminant analysis and longitudinal item response theory models as useful techniques for identifying a threshold for *within-individual* change to identify groups of responders and non-responders. However, regarding estimates of mean score change within an 'improved' anchor group, we maintain that they do not target the location of a threshold and are therefore theoretically biased estimators of within-individual change thresholds (4). Instead, mean change within an 'improved' anchor group has been proposed as more appropriate to guide thresholds for *within-group* changes over time (40,41). Similarly, calculating the difference in mean change in scores between an 'improved' and 'stable' anchor group is not a theoretically appropriate estimator of a within-individual change threshold (38), but instead has been proposed as more suited to between-group differences in change over time (41,42). However, simulations presented within this special section (28) suggest that deviations from normally distributed score changes may pose a challenge to these theoretical ideals. Further planned simulations should help to confirm this (43).

A second observation is that current methods for within-individual thresholds and their clinical application use estimators relying on between-individual variability (4,7). For example, meaningful change threshold estimation typically compares between-individual variation in an anchor measure with between-individual variation in change between two assessment points. And distribution-based indices are based on between-individual variance (e.g., standard deviation of a test score multiplied by a constant representing the level of accepted uncertainty and another variable such as the reliability coefficient). If researchers or clinicians are interested in understanding how a group of patients is

classified over the course of time (and not making a statement about individual patients), then using measures that are based on between-individual variance is likely an appropriate approach (4). However, if a statement about an individual patient is the goal, then we know that between-individual variability is not always a good or justifiable proxy for within-individual variability (29,44–47). In such a situation, the use of within-individual methods (e.g., EMA or related methods to explore intra-individual variation (18)) might be more appropriate. In the call for papers, we encouraged authors to explicitly justify whether thresholds were intended for between-group, within-group, or within-individual interpretations and why it was appropriate to do so. This has led to calls for more nuance in interpretation (7); to pragmatic responses that within-individual change methodology faces challenges in practical applications ((5); but see (18,29) for contrasting examples); to detailed statements on how to interpret a given index and when and where it is appropriate to use (4); as well as wider discussions and explanations of the methods leading to such indices (19,31,35). We especially see the development of appropriate within-individual methods for the identification of change as a key priority that also aligns with current technological developments for practice.

A third point is that in many submissions the variability or uncertainty associated with either the threshold or the change estimate is an important element in interpretation. Knowing the uncertainty associated with a threshold estimate is important, but not always explained or provided. Regardless of the type of variability used and whether a threshold based on meaning to patients or distributions is sought, recognizing and making transparent that there is uncertainty associated with these thresholds is a valuable reminder that none of the methods discussed in this special section offer absolute results. Because the use of meaningful change methodology and distribution-based thresholds has been ritualized to a degree, it is not always considered whether a particular method to determine thresholds is the most appropriate one for a given context. Additionally, emerging mixed methods research relies on classifying particular patients as "changed" for identification in case studies, with limited or no allowance for measurement error, as well as assuming that the classification threshold applies to this particular patient (48,49). Transparency about uncertainty in thresholds and classifications as well as whether it is appropriate to apply a threshold for group or individual change is therefore a key

consideration for developing mixed-methods research agendas around how health outcome measures are used by patients more broadly (50–53). We think that this intersection between epistemology, psychometrics, and various fields of clinical practice contains one of the strongest development opportunities for our understanding of (subjective) health outcome measurement, but substantial work is needed to align theories and practices for a coordinated research effort in this area.

The call for papers was issued to invite discussion, development, as well as state-of-the-art research and practice. We are grateful for the excellent range of submissions received and to all authors and reviewers involved in selecting the published papers, which represent a two-year collective effort. We hope that readers find these papers useful both in developing their own research, but also to help the field to further extend its efforts around patient-centeredness. When we can all agree on what a meaningful change is and how to measure it for a particular patient, measure, and population, then we will have the opportunity to bring about meaningful change in clinical practice and at the social and policy level.

Acknowledgement

This paper is available as a preprint at https://psyarxiv.com/75a8c

Preprint DOI: 10.31234/osf.io/75a8c

REFERENCES

1. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Quality of Life Research. 2013 Oct 1;22(8):1889–905.

2. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S, the Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL). Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res. 2013 Apr;22(3):475–83.

3. King MT, Dueck AC, Revicki DA. Can Methods Developed for Interpreting Group-level Patient-reported Outcome Data be Applied to Individual Patient Management? Med Care. 2019 May;57 Suppl 5 Suppl 1:S38–45.

4. Terwee CB, Peipert JD, Chapman R, Lai JS, Terluin B, Cella D, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res. 2021 Oct;30(10):2729–54.

5. Hays RD, Peipert JD. Between-group minimally important change versus individual treatment responders. Qual Life Res. 20210615th ed. 2021 Oct;30(10):2765–72.

6. Musoro ZJ, Hamel JF, Ediebah DE, Cocks K, King MT, Groenvold M, et al. Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-analysis protocol. BMJ Open. 2018 Jan;8(1):e019117.

7. Trigg A, Griffiths P. Triangulation of multiple meaningful change thresholds for patient-reported outcome scores. Qual Life Res. 2021 Oct;30(10):2755–64.

8. Staunton H, Willgoss T, Nelsen L, Burbridge C, Sully K, Rofail D, et al. An overview of using qualitative techniques to explore and define estimates of clinically important change on clinical outcome assessments. J Patient Rep Outcomes. 2019 Dec;3(1):16.

9. Cocks K, King MT, Velikova G, de Castro G, Martyn St-James M, Fayers PM, et al. Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. European Journal of Cancer. 2012 Jul;48(11):1713–21.

10. Cocks K, Buchanan J. How scoring limits the usability of minimal important differences (MIDs) as responder definition (RD): an exemplary demonstration using EORTC QLQ-C30 subscales. Qual Life Res [Internet]. 2022 Jul 9 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03181-4

11. Wang Y, Devji T, Qasim A, Hao Q, Wong V, Bhatt M, et al. A systematic survey identified methodological issues in studies estimating anchor-based minimal important differences in patient-reported outcomes. Journal of Clinical Epidemiology. 2022 Feb;142:144–51.

12. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Review of Pharmacoeconomics & Outcomes Research. 2011 Apr;11(2):171–84.

13. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Controlled Clinical Trials. 1989 Dec;10(4):407–15.

14. SISAQOL-IMI | IMI Innovative Medicines Initiative [Internet]. [cited 2023 Mar 24]. Available from: https://www.imi.europa.eu/projects-results/project-factsheets/sisaqol-imi

15. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006 Dec;4(1):54.

16. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health and Quality of Life Outcomes. 2006;4(1):79.

17. Blampied NM. Reliable change and the reliable change index: still useful after all these years? tCBT. 2022;15:e50.

18. Smit AC, Snippe E, Bringmann LF, Hoenders HJR, Wichers M. Transitions in depression: if, how, and when depressive symptoms return during and after discontinuing antidepressants. Qual Life Res [Internet]. 2022 Nov 23 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03301-0

19. Bjorner JB, Terluin B, Trigg A, Hu J, Brady KJS, Griffiths P. Establishing thresholds for meaningful within-individual change using longitudinal item response theory. Qual Life Res [Internet]. 2022 Jul 23 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03172-5

20. Li Y. Inferring meaningful change in quality of life with posterior predictive distribution: an alternative to standard error of measurement. Qual Life Res [Internet]. 2022 Sep 9 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03239-3

21. Wyrwich KW, Kitchen H, Knight S, Aldhouse NVJ, Macey J, Mesinkovska N, et al. Using qualitative methods to establish the clinically meaningful threshold for treatment success in alopecia areata. Qual Life Res [Internet]. 2022 Jul 12 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03170-7

22. Bartlett SJ, Bykerk VP, Schieir O, Valois MF, Pope JE, Boire G, et al. "From Where I Stand": using multiple anchors yields different benchmarks for meaningful improvement and worsening in the rheumatoid arthritis flare questionnaire (RA-FQ). Qual Life Res [Internet]. 2022 Sep 8 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03227-7

23. Jones SMW, Gaffney A, Unger JM. Common methods of determining meaningful change in clinical practice: implications for precision patient-reported outcomes. Qual Life Res [Internet]. 2022 Sep 10 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03246-4

24. Wyrwich KW, Norman GR. The challenges inherent with anchor-based approaches to the interpretation of important change in clinical outcome assessments. Qual Life Res [Internet]. 2022 Nov 18 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03297-7

25. Griffiths P, Sims J, Williams A, Williamson N, Cella D, Brohan E, et al. How strong should my anchor be for estimating group and individual level meaningful change? A simulation study assessing anchor correlation strength and the impact of sample size, distribution of change scores and methodology on establishing a true meaningful change threshold. Qual Life Res [Internet]. 2022 Nov 19 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03286-w

403 26. Griffiths P, Terluin B, Trigg A, Schuller W, Bjorner JB. A confirmatory factor analysis approach
404 was found to accurately estimate the reliability of transition ratings. Journal of Clinical
405 Epidemiology. 2022 Jan;141:36–45.

406 27. Terluin B, Koopman JE, Hoogendam L, Griffiths P, Terwee CB, Bjorner JB. Estimating
407 meaningful thresholds for multi-item questionnaires using item response theory. Qual Life Res
408 [Internet]. 2023 Feb 13 [cited 2023 Feb 24]; Available from:
409 https://link.springer.com/10.1007/s11136-023-03355-8

410 28. Qin S, Nelson L, Williams N, Williams V, Bender R, McLeod L. Comparison of anchor-based
411 methods for estimating thresholds of meaningful within-patient change using simulated PROMIS
412 PF 20a data under various joint distribution characteristic conditions. Qual Life Res [Internet].
413 2022 Nov 13 [cited 2023 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-
414 03285-x

415 29. Bringmann LF, van der Veen DC, Wichers M, Riese H, Stulp G. ESMvis: a tool for visualizing
416 individual Experience Sampling Method (ESM) data. Qual Life Res [Internet]. 2020/11/23 ed.
417 2020 Nov 22; Available from: https://www.ncbi.nlm.nih.gov/pubmed/33222049

418 30. Poon JL, Marshall C, Johnson C, Pegram HC, Hunter M, Kan H, et al. A qualitative study to
419 examine meaningful change in physical function associated with weight-loss. Qual Life Res
420 [Internet]. 2022 Jul 22 [cited 2023 Feb 24]; Available from:
421 https://link.springer.com/10.1007/s11136-022-03191-2

422 31. Peipert JD, Hays RD, Cella D. Likely change indexes improve estimates of individual change on
423 patient-reported outcomes. Qual Life Res [Internet]. 2022 Aug 3 [cited 2023 Feb 24]; Available
424 from: https://link.springer.com/10.1007/s11136-022-03200-4

425 32. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest
426 real difference, a link between reproducibility and responsiveness. Quality of Life Research.
427 2001;10(7):571–8.

428 33. Lee MK, Peipert JD, Cella D, Yost KJ, Eton DT, Novotny PJ, et al. Identifying meaningful
429 change on PROMIS short forms in cancer patients: a comparison of item response theory and
430 classic test theory frameworks. Qual Life Res [Internet]. 2022 Sep 24 [cited 2023 Feb 24];
431 Available from: https://link.springer.com/10.1007/s11136-022-03255-3

432 34. Ho EH, Verkuilen J, Fischer F. Measuring individual true change with PROMIS using IRT-based
433 plausible values. Qual Life Res [Internet]. 2022 Oct 25 [cited 2023 Feb 24]; Available from:
434 https://link.springer.com/10.1007/s11136-022-03264-2

435 35. Andrae DA, Foster B, Peipert JD. Comparison of raw and regression approaches to capturing
436 change on patient-reported outcome measures. Qual Life Res [Internet]. 2022 Sep 22 [cited 2023
437 Feb 24]; Available from: https://link.springer.com/10.1007/s11136-022-03196-x

438 36. Cronbach LJ, Furby L. How we should measure 'change': Or should we? Psychological Bulletin.
439 1970;74(1):68–80.

440 37. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley;
441 1968.

442 38. Vanier A, Sébille V, Blanchin M, Hardouin JB. The minimal perceived change: a formal model
443 of the responder definition according to the patient's meaning of change for patient-reported
444 outcome data analysis and interpretation. BMC Med Res Methodol. 2021 Dec;21(1):128.

39. Gerlinger C, Schmelter T. Determining the non-inferiority margin for patient reported outcomes: Determining the non-inferiority margin for patient reported outcomes. Pharmaceut Statist. 2011 Sep;10(5):410–3.

40. Musoro JZ, Coens C, Fiteni F, Katarzyna P, Cardoso F, Russell NS, et al. Minimally Important Differences for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer. JNCI Cancer Spectrum. 2019 Sep 1;3(3):pkz037.

41. Sabah SA, Alvand A, Beard DJ, Price AJ. Minimal important changes and differences were estimated for Oxford hip and knee scores following primary and revision arthroplasty. Journal of Clinical Epidemiology. 2022 Mar;143:159–68.

42. Bell ML, Dhillon HM, Bray VJ, Vardy JL. Important differences and meaningful changes for the Functional Assessment of Cancer Therapy-Cognitive Function (FACT-Cog). J Patient Rep Outcomes. 2018 Dec;2(1):48.

43. Vanier A, Leroy M, Hardouin JB. 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. Methods. 2022 Aug;204:396–409.

44. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proceedings of the National Academy of Sciences. 2018 Jul 3;115(27):E6106–15.

45. Harvill LM. An NCME Instructional Module on. Standard Error of Measurement. Educational Measure: Issues Practice. 1991 Jun;10(2):33–41.

46. McAleavey AA. When (Not) to Rely on the Reliable Change Index [Internet]. Open Science Framework; 2021 Nov [cited 2023 Feb 25]. Available from: https://osf.io/3kthg

47. Molenaar P, Campbell CG. The new person-specific paradigm in psychology. Current Directions in Psychological Science. 2009;18(2):112–7.

48. De Smet M, Acke E, Cornelis S, Truijens F, Notaerts L, Reitske Meganck, et al. Understanding 'patient deterioration' in psychotherapy from patients' perspectives: A mixed methods multiple case study. 2022 [cited 2023 Feb 25]; Available from: https://rgdoi.net/10.13140/RG.2.2.17796.60802

49. Desmet M, Van Nieuwenhove K, De Smet M, Meganck R, Deeren B, Van Huele I, et al. What too strict a method obscures about the validity of outcome measures. Psychother Res. 20210204th ed. 2021 Sep;31(7):882–94.

50. McClimans L. Interpretability, validity, and the minimum important difference. Theor Med Bioeth. 2011/07/28 ed. 2011 Dec;32(6):389–401.

51. McClimans LM. First person epidemiological measures: vehicles for patient centered care. Synthese. 2021 May;198(S10):2521–37.

52. Truijens FL, Desmet M, De Coster E, Uyttenhove H, Deeren B, Meganck R. When quantitative measures become a qualitative storybook: A phenomenological case analysis of validity and performativity of questionnaire administration in psychotherapy research. Qualitative Research in Psychology. 2019;19(1):244–87.

53. Truijens FL, Van Nieuwenhove K, De Smet MM, Desmet M, Meganck R. How questionnaires shape experienced symptoms. A qualitative case comparison study of questionnaire

administration in psychotherapy research. Qualitative Research in Psychology. 2021;19(3):806–30.

Table 1: Classification of meaningful change thresholds

Dimension	Options
1: Level of interpretation	A: Group B: Individual
2: Type of comparison	A: Difference (cross-sectional) B: Change over time* C: Difference in change over time
3: Magnitude	A: Minimal B: Larger than minimal (e.g., moderate or large)

Note. *May be further split according to improvement / worsening.

Table 2: Characteristics of included papers.

Paper	Meaningful change or Distribution-based	Level of interpretation	Type of comparison	Magnitude
Andrae et al. (35)	Distribution-based	Individual	Change over time	N/A
Bartlett et al. (22)	Meaningful change	Group	Change over time	Minimal and larger than minimal
Bjorner et al. (19)	Meaningful change	Individual	Change over time	Minimal
Cocks & Buchanan (10)	N/A	Individual	Change over time	N/A
Griffiths et al. (25)	Meaningful change	Group, Individual	Change over time	Minimal
Ho et al. (34)	Distribution-based	Individual, Group	Change over time	N/A
Jones et al. (23)	Meaningful change	Individual	Change over time	Not specified
Lee et al. (33)	Both	Individual	Change over time	Minimal
Li et al. (20)	Distribution-based	Individual	Change over time	N/A
Peipert et al. (31)	Distribution-based	Individual	Change over time	N/A
Poon et al. (30)	Meaningful change	Individual	Change over time (hypothetical)	Minimal
Qin et al. (28)	Meaningful change	Individual	Change over time	Not specified
Smit et al. (18)	Both	Individual	Change over time	Meaningful[a]
Wyrwich & Norman (24)	Meaningful change	General	General	General
Wyrwich et al. (21)	Meaningful change	Individual	Change over time (hypothetical)	Meaningful[b]

Note. [a] 'meaningful' was defined as the patient clearly noticing a change in daily life and/or experiencing discomfort as a result of the change.

[b] 'meaningful' was defined as the amount of change that patients considered a (hypothetical) treatment success.

506

507

508