Title

Commentary on “When (Not) to Rely on the Reliable Change Index”

Journal

Clinical Psychology Science and Practice, 31(3)

ISSN

0969-5893

Authors

Hays, Ron D

Reise, Steven P

Publication Date

2024

DOI

10.1037/cps0000210

Peer reviewed

Clinical Psychology: Science and Practice

ISSN: 0969-5893

2024, Vol. 31, No. 3, 367–369
https://doi.org/10.1037/cps0000210

COMMENTARY

Commentary on "When (Not) to Rely on the Reliable Change Index"

Ron D. Hays[1] and Steven P. Reise[2]
[1] Division of General Internal Medicine & Health Services Research, Department of Medicine, UCLA
[2] Department of Psychology, UCLA

McAleavey (2024) provides extensive criticism of the reliable change index (RCI) and some general information about alternatives, but we have concerns about the article. This commentary briefly mentions seven areas of concern: (a) minimally important difference (MID) estimates are not appropriate for assessing individual change; (b) analytic choices are not either/or; (c) test–retest reliability is not superior to internal consistency reliability; (d) coefficient of repeatability versus the instrument's RCI; (e) sum scores; (f) feasibility of obtaining multiple observations per individual; and (g) alternatives to the RCI.

MID Estimates Are Not Appropriate for Individual Change

The MID is an estimate of the threshold for the size of a mean difference on a measure that is minimally important. The minimally important change (MIC) is the MID that focuses on mean change on a measure. The MIC is a subset of responsiveness to change intended to represent the smallest mean difference that is important enough to care about. MIC estimates are motivated by the fact that trivial group mean change can be statistically significant if the sample size is large enough.

McAleavey claims that the RCI "has been considered a 'distribution-based MID'" because it is based on population statistics. There is no such thing as population statistics. There are sample statistics and population parameters. So-called distribution-based MID estimates such as 0.5 *SD* or the standard error of measurement do not provide estimates of the size of change that is minimally important. Moreover, while he acknowledges that the RCI and MID "are different quantities," and that the MID estimates are usually smaller than the coefficient of repeatability, the MID is used for interpreting group mean differences rather than individual change (Hays & Peipert, 2021).

Analytic Choices Are Not Either/Or

Data analysis choices are often not either this or that. McAleavey expresses concern throughout the article about how large individual change needs to be for it to reach the conventional $p < .05$ significance level when using the RCI. To his credit, McAleavey refers to Donaldson's (2008) suggestion to consider other levels of statistical significance given how conservative the RCI at $p < .05$ is in detecting true change. Indeed, various researchers have now advocated for potentially using less stringent *p*-levels for the RCI to be consistent with change perceived to be important by patients (Peipert et al., 2023).

McAleavey notes that end-state functioning may be more clinically relevant than the difference score and notes that both change, and end-state can be important. Indeed, Jacobson and Truax (1991) suggested using both significant and clinically meaningful change together. For example, 37% of the adults with low back pain in a randomized trial improved significantly on the Impact Stratification Scale (ISS) over these 6 weeks, and 59% reported on a retrospective change item that they were better (Hays & Peipert, 2021). Among those who improved significantly on the ISS, 89% reported they were better on the retrospective rating item. Thirty-three percent of the sample improved significantly and reported improvement on the retrospective change item (statistically and clinically meaningful), 4% improved significantly but did not report that they were better on the retrospective change item (statistically but not clinically meaningful), 26% did not improve significantly but reported improvement on the change item, and 37% did not improve significantly or report improvement on the change item. This example illustrates how both statistical significance and perceived change provide important information together.

Test–Retest Reliability Is Not Superior to Internal Consistency Reliability

McAleavey advocates for the use of test–retest reliability rather than internal consistency reliability for the RCI. However, practical problems with obtaining a good test–retest estimate are not emphasized enough. Picking the optimal time interval for test–retest reliability is difficult. It should not be too soon such that responses at

Ron D. Hays https://orcid.org/0000-0001-6697-907X
Ron D. Hays was supported by the Center for Health Improvement of Minority Elderly (CHIME), Resource Centers for Minority Aging Research (RCMAR) under the National Institute of Health/National Institute of Aging (NIH/NIA) Grant P30-AG021684. The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Ron D. Hays served as lead for conceptualization, writing–original draft, and writing–review and editing. Steven P. Reise served in a supporting role for writing–review and editing.

Correspondence concerning this article should be addressed to Ron D. Hays, Division of General Internal Medicine & Health Services Research, Department of Medicine, UCLA, 1100 Glendon Avenue, Suite 850, Los Angeles, CA 90024, United States. Email: drhays@ucla.edu

the second assessment are simply memories of the first assessment, yet not so long that true change in the construct has occurred during the time interval between the initial and subsequent assessment. Importantly, test–retest underestimates reliability when there is true underlying change. As Reeve et al. (2007) noted:

> ISOQOL respondents agreed that as a minimum standard a multi-item PRO measure should be assessed for internal consistency reliability. … However, they did not support as a minimum standard that a multi-item PRO measure should be required to have evidence of test–retest reliability. They noted practical concerns regarding test–retest reliability; primarily that some populations studied in PCOR are not stable and that their HRQOL can fluctuate. This phenomenon would reduce estimates of test–retest reliability, making the PRO measure look unreliable when it may be accurately detecting changes over time. In addition, memory effects will positively influence the test–retest reliability when the two survey points are scheduled close to each other (p. 1895).

Coefficient of Repeatability Versus the Instrument's RCI

The article states that an instrument's RCI is indicated by the amount of individual change needed to be significant at $p < .05$. We think it is confusing to refer to the RCI formula as the "RC index" and the amount of change needed to be significant for an instrument as the instrument's RCI because an instrument does not have a single reliability or *SD*. The amount of change needed for significance at a given *p*-value, equivalent to the RCI, has been widely referred to as the coefficient of repeatability.

Sum Scores

McAleavey states that if "two identical scores may come from two meaningfully different item response combination, sum scores are less valid" and "two patients with equal scores might report meaningfully different experiences" (p. 7). If a measure is unidimensional, the sum score is a sufficient statistic for the Rasch measurement model. And even when items are differentially weighted (e.g., graded response model), sum scores and item response theory scores are very highly correlated (Embretson & Reise, 2013).

Feasibility of Collecting Multiple Observations Per Individual

McAleavey suggests getting more than two observations whenever possible and using all data in the analysis (p. 16). We doubt that there are objections to his conclusion that: "Researchers should collect more than two observations when they hope to examine change with fidelity, and clinicians should consider all available data when making determinations of change." However, collecting enough observations to obtain an accurate estimate of individual-level variation is often extremely challenging. Individual-level variation and change can be estimated using time series data, but many observations are needed for accurate estimation—for example, simulation modeling analyses require a minimum of 10 observations (Borckardt et al., 2008). Smit et al. (2023) analyzed data from 56 respondents who provided a median of 27 weekly depressive symptom assessments during and after antidepressant discontinuation, but this represented only 26% of those who were eligible for the study. The rest were excluded due to refusing to participate, missing data, and lost to follow-up.

Alternatives to the RCI

The article is unclear about what approach should be used instead of the RCI when there are only two data points. McAleavey suggests that the RCI is the simplest method but least desirable compared to measurement error models such as structural equation modeling, but how individual change is estimated in these alternatives is not provided. McAleavey also mentions that a multiple timepoint extension of the RCI that uses linear regression with measurement error is available (McAleavey, 2022), but the article is devoid of information about how significant individual change is evaluated.

There are also limited specifics about analyzing data when there are a large number of data points for individuals. Borckardt et al. (2008) provide explicit information about how this can be done using simulation modeling that accounts for autocorrelations in the data stream.

Summary

McAleavey notes flaws with estimating individual change with two data points using the RCI. Although better alternatives are mentioned, there are limited specifics about implementing them. Instead, there are statements about alternatives such as:

> Neither is universally beneficial and both have been criticized.
>
> These methods will, however, generally be even less sensitive than the RCI, since they do not treat small true changes as meaningful.
>
> This method entails many assumptions, but it may be relevant when speaking about groups.
>
> IRT is not a perfect solution for the RCI.
>
> While no specific alternative will be advisable in every case.

The caveats about alternatives are important, but this makes clear that there are no simple solutions to the problems raised. It is better to have more data and to go beyond the RCI in assessing individual change. The main takeaway message readers may have from the McAleavey article is that the RCI is not good but it is not clear what alternative should be used.

References

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time series analysis. *American Psychologist*, *63*(2), 77–95. https://doi.org/10.1037/0003-066X.63.2.77

Donaldson, G. (2008). Patient-reported outcomes and the mandate of measurement. *Quality of Life Research*, *17*(10), 1303–1313. https://doi.org/10.1007/s11136-008-9408-4

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Hays, R. D., & Peipert, J. D. (2021). Between-group minimally important change versus individual treatment responders. *Quality of Life Research*, *30*(10), 2765–2772. https://doi.org/10.1007/s11136-021-02897-z

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

McAleavey, A. A. (2022). *ReliableTrendIndex: Utilities for (more reliable) reliable change analysis*. Retrieved February 28, 2024, from https://github.com/andrewmcaleavey/ReliableTrendIndex

McAleavey, A. A. (2024). When (not) to rely on the reliable change index: A critical appraisal and alternatives to consider in clinical psychology. *Clinical Psychology: Science and Practice*, *31*(3), 351–366. https://doi.org/10.1037/cps0000203

Peipert, J. D., Hays, R. D., & Cella, D. (2023). Likely change indexes improve estimates of individual change on patient-reported outcomes. *Quality of Life Research*, *32*(5), 1341–1352. https://doi.org/10.1007/s11136-022-03200-4

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., Cella, D., & PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Smit, A. C., Snippe, E., Bringmann, L. F., Hoenders, H. J. R., & Wichers, M. (2023). Transitions in depression: If, how, and when depressive symptoms return during and after discontinuing antidepressants. *Quality of Life Research*, *32*(5), 1295–1306. https://doi.org/10.1007/s11136-022-03301-0

Received March 3, 2024
Accepted March 7, 2024 ■