The Duration-adjusted Reliable Change Index (DaRCI): defining clinically relevant symptom changes of varying durations

Marieke A. Helmich

University of Groningen, University Medical Center Groningen, Interdisciplinary Center Psychopathology and Emotion Regulation, The Netherlands

Author Note

Correspondence concerning this article should be addressed to Marieke A. Helmich, University of Groningen, University Medical Center Groningen, Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), P.O. Box 30.001 (CC72), 9700 RB Groningen, the Netherlands. E-mail: m.a.helmich@umcg.nl

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-CoG-2015; No 681466 to M. Wichers).

Abstract

Identifying relevant symptom shifts as they enfold can be challenging, as the time period over which they take place is not uniform for all people. This paper proposes an adaptation of the well-established Reliable Change Index (RCI) that allows researchers and clinicians to explore the presence of symptom changes of varying durations in individual patients' time series: the Duration-adjusted RCI (DaRCI). The DaRCI takes the RCI cut-off score for change between two points as a starting point, and proportionally extends this measure over multiple observations, while maintaining reliability at a given confidence level. Researchers must choose the relevant time period between two observations, and additional increments are added accordingly.

To illustrate the ability of this method to detect changes of various durations, simulated depressive symptom time series with varying degrees of discontinuity and overall mean change in scores were used. The results show that the DaRCI thresholds over two, three and four observations were effective at identifying the simulated change periods over multiple time points, starting from relatively gradual change slopes (picking up reliable changes in 20-60% of simulated time series if the overall change was large enough), to highly discontinuous changes (up to 100% accuracy).

The DaRCI may be particularly useful for identifying shifts in symptoms that appear relatively abrupt, which can help indicate when a patient is showing significant improvement or deterioration. Its ease of use makes it suitable for application in the clinical context, and is a promising method to explore different change durations in clinical populations.

Keywords: duration of change; within-person; reliable symptom change; repeated measurement; idiographic method;

Background

Identifying if and when a patient's psychological symptoms have changed in a clinically relevant way is an integral part of many treatment settings, and studies of therapeutic interventions. Typically, methods to determine clinical change identify when a patient reaches a score over or under a certain threshold (e.g., within the range of a non-clinical population norm score [1]), or shows a change in scores that meets a cut-off (e.g., 50% reduction [2]) or combination of criteria (e.g., minimal score reduction and statistical significance [3]). The Reliable Change Index (RCI) is a widespread method that determines whether the variability in a person's measurements on a symptom questionnaire are more likely to be due to the instrument's precision (measurement error), or due to an actual clinical change [1,4–6]. Determining if a symptom change indicated by a questionnaire is reliable is important to establish whether a drop or increase in scores is not merely due to chance, and can aid decisions on whether to start, continue, end, or alter the intensity of treatment [7].

What most reliable change methods do not incorporate, however, is the time frame over which a change occurred. Some methods focus on pre- to post-treatment change –which can take weeks or months–, where other methods aim to identify symptom shifts as they occur between or even within therapy sessions [8–10]. For instance, in research on depression and anxiety, "sudden gains" and "sudden losses" in symptoms are identified as changes between therapy sessions that combine a pre-defined minimum magnitude of change with the requirement that it takes place within a short period of time, most often a week [11–14]. Yet, symptom reductions of 50% or more over three to four weeks of treatment have also been described as "rapid" in studies of early response, and are certainly considered clinically relevant [15,16].

Differences in the duration of change should be taken into account when identifying a period of significant improvement, as the time it takes to change can be clinically meaningful in itself [17,18]. Imagine two individuals who show the same reliable reduction in scores (e.g., –15 points). For person A the cut-off is met relatively suddenly, from one week to the next, while person B shows slower, gradual improvement and meets the same reduction only after three weeks. These individuals both show a reliable improvement according to this criterion, but due to the different timings, the process

appears as a sudden change for one (person A) and a gradual improvement for the other (person B). Here, using a single criterion is effective at detecting a minimal reduction for both people, but it does nothing to capture the difference in "velocity" of the changes.

Now compare the changes shown by those two individuals to those of a third person. If person C showed a pronounced decline over multiple weeks (e.g., steps of –12, –9, –11 points), this pattern meets the minimal change criterion of –15 points after the second step but (–21 compared to the first point), unlike person A, the changes between two adjacent points (over one week) never meet the minimal change criterion. Yet, the overall change shown by person C is larger overall (–32) and may also be considered a clinically relevant and "rapid" improvement, despite the longer time frame (compared to person A) (cf. [15,16]). Using a single cut-off that is unadjusted for the duration of a symptom change allows one to determine reliable change over one interval (person A), but runs the risk of missing smaller within-week changes that continue over multiple weeks and culminate into a reliable change over a longer period (person C). Thus, if different durations could be accounted for by a change criterion that requires that longer changes must also be larger as a whole, person C's consistent improvement may be identified as a reliable change that is similar in its clinical relevance to the rapid one-week change shown by person A. In short, testing one criterion for changes with no regard for time, or considering only one shifts of one specific duration (e.g., one week, as illustrated above) disregards clinically relevant changes that occur over consecutive time points.

As routinely measuring (former) patients' psychological complaints and collecting time series of repeated assessments within individuals has become common practice in the context of therapy and relapse prevention [19–24], studies mapping repeated symptom assessments have shown that psychopathological change is often characterized by nonlinearity and abrupt changes [25–28]. Clinically, this is relevant as particularly sudden shifts may indicate that the patient may have experienced a 'transition' to a better or worsened state, which could be predictive of their treatment outcomes [11,13,14,28–30]. Sudden gains (symptom improvements), for instance, may occur when a therapy session or intervention has been especially effective [14,31–33], while sudden losses (deteriorations) indicate when a patient is less likely to benefit from treatment, and should be identified

as soon as possible to prevent the treatment from failing [10,34]. A pattern of steady early improvement over the first few treatment sessions has also been linked to better treatment outcomes [16,31,32,35,36], and conversely, early changes that did not conform to the expected response patterns have been linked to poorer outcomes [37]. Thus, change patterns of varying duration and magnitude have been described in the psychotherapy context [38–40], but flexible methods to detect shifts, as they unfold, and over different durations, are lacking.

To summarize, various methods exist to determine whether a relevant change in symptoms has occurred at the within-person level, but these typically make no particular assumption about the time it took for the symptoms to change. Furthermore, even those that examine sudden gains and losses in repeated assessment data do not provide solutions to identify abrupt changes that extend over multiple time points or therapy sessions. More data-intensive methods may focus on testing the significance of an overall symptom change with a regression model [6,41–43], or try to identify abrupt shifts with a change-point model [44,45], but these methods may be difficult to implement in real-time in the course of clinical practice, as they require more data than may be available in early stages of treatment, and also a fair level of statistical knowledge to be conducted (cf. [46]). Thus, adjusting a simple method like the RCI may help to identify both reliable symptom changes that are more sudden as well as reliable slower gradual symptom changes over time.

In this brief report, I propose a method, based on the well-established Reliable Change Index [4], which allows researchers and clinicians to explore the presence of symptom changes of varying durations in individual patients' time series: the Duration-adjusted Reliable Change Index (DaRCI). A simulation study is conducted to test the DaRCI's ability to pick up periods of relevant change, and discontinuous change in particular, in the context of a larger overall symptom time series.

Method

Material

This study used the Dutch Symptom Checklist-90 (SCL-90) depression subscale [47,48] as a basis for our illustration of the DaRCI method. This questionnaire consists of sixteen items that ask to what extent one was bothered by particular depressive symptoms (e.g., "feeling blue") on a five-point scale ranging from "not at all" to "very much".

Reliable Change Index

The RCI was developed as a method to ensure that any identified pre- to post-treatment change was a reliable change that could be distinguished from measurement error [1,4]. The RCI can be used to calculate a threshold at which the difference between a pre- and post-measurement for one person is, with a 95% two-tailed confidence level, "unlikely to occur without actual change" (Ref. [4], p. 14). It uses the standard error of measurement (SE_{m}) from a population norm of a given instrument to calculate the minimum score necessary to exceed a change that could be due to measurement inaccuracy (SE_{diff}, standard error of difference). Note that the RCI uses between-persons information for the standard error of measurement[1] regarding the spread of the distribution of test-retest reliability given no change, and uses this to test whether within-person changes are reliable. It is defined as:

$$RCI = SE_{\mathrm{diff}} \times Z,$$

$$SE_{\mathrm{diff}} = \sqrt{2(SE_{\mathrm{m}})^2}.$$

For this study, $SE_{\mathrm{m}} = 4.37$; taken from the Dutch SCL-90 depression subscale, based on the psychiatric outpatient norm group of 5,621 patients (cf. SCL-90 manual; Ref. [47]). Thus, $SE_{\mathrm{diff}} = \sqrt{2(4.37)^2} = 6.18$, and $RCI_{95} = 6.18 \times 1.96 = 12.113$. Where 1.96 is the *Z*-score for a 95% confidence level, and a score of ~13 is a reliable change between two observations on the SCL-90

[1] Ideally, the same time frame is used for the calculation of the standard error of measurement (based on the test-retest reliability [6,41]) as for the investigated difference score. However, in practice, this time frame is usually not considered and the RCI is calculated with the available information and applied to pre-posttest scores.

depression subscale. The score is rounded up to ensure the confidence level is maintained, as the (difference) scores are always integers, and obtaining a score of 12.113 is not possible.

Duration-adjusted RCI (DaRCI)

The DaRCI is proposed as an adaptation of the RCI, with the aim to capture symptom changes of varying durations, particularly changes that appear as sudden or large in overall scope. The DaRCI requires setting a fixed time period between two points (e.g., one week) as a basis for extension when more points are added (i.e., when testing change over longer durations), and allows one to calculate thresholds for each additional increment of time (e.g., each added week). Like the regular RCI, the DaRCI tests the difference score between two points, but it accounts for instances where the two compared points are farther apart by proportionally increasing the change threshold.

To detect reliable change from a given starting point (t_{start}) to the last observation in a chosen period (t_n) while considering the additional time between these observations, the original RCI threshold (based on *n* observations = 2) is divided by 2, and multiplied by the number of observations in the range of interest. To calculate the DaRCI critical change threshold for a particular *Z*-value (confidence level) and number of observations (*n*):

$$DaRCI = \left(\frac{SE_{\text{diff}}}{2}\right) \times Z \times n$$

Essentially, the RCI threshold is reduced to a range of uncertainty around a single point, and then proportionally extended for the number of observations (i.e., period of time) at hand, while maintaining the chosen confidence level (e.g., 95%). In doing so the DaRCI provides a way for researchers to detect symptom changes over various durations with the same degree of reliability, even if some shifts take longer. Applied to our illustrative sample and instrument, with the SE_{diff} for the SCL-90 depression subscale incorporated: $DaRCI = (6.18/2) \times Z \times n$. For 95% confidence, this can be rewritten as: $DaRCI_{95} = 3.09 \times 1.96 \times n = 6.0564n$. Where 6.0564 is the aforementioned range of uncertainty around one point.

Alternatively, we can calculate the *Z*-score for a particular change over time (i.e., the difference between two assessments (Δy), divided by the number of observations (*n*) in the range of interest. This can be useful to compare the relative magnitude of multiple identified changes. In formula form:

$$DaRCI_Z = \frac{\left(2\frac{\Delta_y}{n}\right)}{SE_{\text{diff}}}.$$

The DaRCI thresholds for different increments are presented in Table 1. Using this method, higher and lower confidence levels may also be calculated and explored, and all DaRCI critical threshold scores are rounded up to maintain the cut-off ≥ Z requirement.

Table 1

The Duration-adjusted Reliable Change Index (DaRCI) over different durations and confidence levels for the SCL-90 depression subscale

	n = 2 e.g., 1 week	*n* = 3 e.g., 2 weeks	*n* = 4 e.g., 3 weeks	*n* = 5 e.g., 4 weeks
90% confidence *Z* ≥1.645	10.17 ≈ 11	15.25 ≈ 16	20.33 ≈ 21	25.42 ≈ 26
95% confidence *Z* ≥ 1.96	12.11 ≈ 13	18.17 ≈ 19	24.23 ≈ 25	30.28 ≈ 31
99% confidence *Z* ≥ 2.58	15.94 ≈ 16	23.92 ≈ 24	31.89 ≈ 32	39.86 ≈ 40

Note. To calculate the cut-off scores, the following formula was used: $DaRCI = \left(\frac{6.18}{2}\right) \times Z \times n$. Where, *n* is the number of observations over which the change occurs. The scores are rounded up to maintain the required minimum Z-score. Change between two observations is equal to the original RCI.

The DaRCI does not prescribe over which time period the change must take place. Instead, the time interval between two observations serves as a basis for the other increments over which it is extended. Researchers must choose the duration of time over which detecting a change would be of interest (e.g., based on the clinical literature, pilot studies, or other conceptual grounding). For instance, if the change between t_{start} and t_2 occurs over one week, then the DaRCI will be proportionally extended for two, three, four weeks, etc. (as the limits of a given scale allow), if t_{start} to t_2 occurs over two weeks, reliable change can be calculated for four, six, eight weeks, etc. It is thus an important theoretical and clinical consideration what durations of "change" are relevant to capture, although the DaRCI lends itself precisely to exploring a number of different durations.

Analysis

To test the accuracy of the DaRCI thresholds for change over two (Tn2), three (Tn3) and four (Tn4) time points, the frequency at which modeled shifts were correctly identified in simulated repeated symptom assessments was examined. A set of 10,000 time series with a length of 15 points was simulated, to reflect a typical duration of psychological treatment [49,50]. Each time series was drawn from a randomized normal distribution with a mean of 0 and variance concurrent with the SCL-90 depression subscale SE_{m} of 4.37. The time series were then fitted to different overall symptom reductions (L = −10, −20, −30, −40), and different degrees of discontinuity. Specifically, five shapes of increasingly abrupt change around the midpoint of the time series were modeled: (a) gradual change as a linear function (no internal shift); (b) a smooth curve formed by a sigmoid function with a slope of 1 at the origin (k), and the midpoint set to point 7.5 (the steepest decline occurs between observation 7 and 8); three mean shifts where change takes place over an increasingly short time period: (c) a step-function with the decrease occurring over four points (n = 4); (d) a step-function with the decrease occurring over three points (n = 3); (e) a step-function with the shift occurring from one point to the next (n = 2).

These change patterns were chosen to explore the interplay between the strength of the overall slope and different degrees of discontinuity. For instance, at an overall reduction of 10 points and a linear function, any identified 'change' would be due to a random fluctuation (as the (Da)RCI thresholds start at 13 points for n = 2). At the other extreme, at −40 points of overall change, we would expect that, even if the discontinuities are more gradual (like in the sigmoid function, or the step change over four points), the DaRCI thresholds will often detect the modeled shifts around the correct time period. Furthermore, three out of the five models are step-functions, which allow certainty about when a shift starts and ends, as this is explicitly modeled (in contrast to the sigmoid curve). Moreover, the durations of those shifts correspond to the different durations of the DaRCI thresholds we tested: over two, three and four points, and should thus provide a good test of the thresholds' sensitivity.

In short, DaRCI_{95} thresholds were calculated for change over two (Tn2), three (Tn3) and four observations (Tn4) and applied to the simulated time series to identify periods over which the criteria

for a reliable change were met. Particularly, we examined the sensitivity of the DaRCI thresholds for finding the modeled discontinuities in the middle of the dataset, and the extent to which they were specific to indicating the intended shift, rather than random fluctuations.

Results

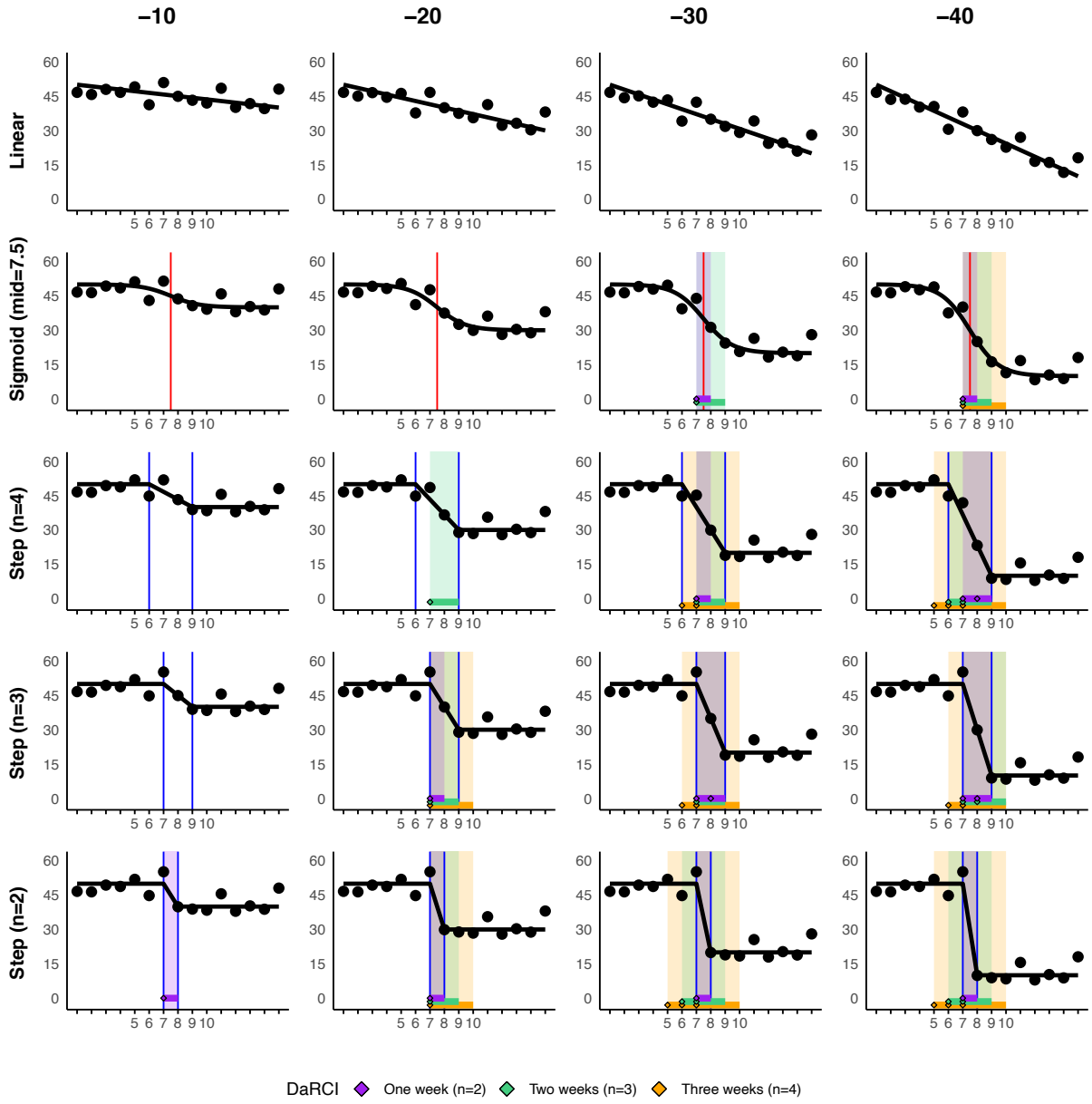
Case illustration

To give a visual illustration of the ability of the DaRCI to identify the intended periods of relevant change, see Figure 1. The indicated changes for the simulated time series in Figure 1 demonstrate clearly how the DaRCI was able to pick up on the periods of increased discontinuity, and thus, relevant change – especially when overall change is at least 30 points. At a lower overall reduction, of 20 points, only the more discontinuous changes were detected in this case example.

Another noteworthy point is that the DaRCI provides thresholds for reliable changes of different duration, but does not itself specify what the most best fitting duration of an identified shift is. Instead, if a score difference meets criteria for the DaRCI at multiple durations, the transitions will simply overlap. This can be seen, for instance, in the bottom row of plots (Step (n=2)) in Figure 1, where change is modeled between two points, but the DaRCI over four points (n=4) also picks up this shift as reliable. While this is correct in that the Tn4 threshold is met, visually it is apparent that the true shift occurs over a shorter duration. It is apparent that the DaRCI over four points cannot identify the exact location of the shift by looking at the multiple starting points on the x-axis. This is in contrast to the DaRCI over two points, which finds change only at one point (between observation 7 and 8). The reverse also happens, where the DaRCI over two points cannot identify the location of a shift in the Step-function over n = 3 and n = 4, and thus places multiple starting points in these longer change periods. However, the relative strength of these changes can be differentiated by their *Z*-scores, with higher scores indicating the best fit for that particular change. For instance, the aforementioned shift (Figure 1, –40, Step (n=2)) has a Z-score of –7.3 when the DaRCI is calculated for the change over two points (Tn2), and –5 at Tn3, and –3.7 for the Tn4 change duration.

Figure 1

Case demonstration of a simulated time series, with increasing levels of overall decline, and increasingly abrupt shifts around the middle of the time series.



Note. Increasing levels of overall score reductions (–10 to –40) are shown on the general X-axis (see headers), and increasing discontinuity around the midpoint on the general Y-axis (from a linear model without a shift, to an abrupt step-change between two points). Periods where reliable changes were indicated by the DaRCI (with durations between one to three weeks) are shown with the colored areas (i.e., purple, green, orange), with the point markers indicating the 'start' of a reliable change. The red vertical line indicates the simulated 'midpoint' of the sigmoid pattern, and the blue vertical lines indicate simulated start and end points for the step function.

Results of the simulation

In Figure 2, the results of the 10,000 simulations are visually represented as a heatmap, with higher values (darker purple) representing a higher frequency of DaRCI-indicated change start points for different durations (Tn2, Tn3 and Tn4).

Linear. Looking at the first row of subplots, for Linear change, we see a low degree of false positives across the different thresholds. There is a slightly higher degree of false positives for the DaRCI threshold for changes between two points (Tn2), indicating that the criterion picked up approximately 5% random fluctuations rather than true changes – this also applied to the DaRCI for change over two points in the other overall change models. The DaRCI criteria for changes over three or four points showed negligible rates of false positives, 0.5% at Tn3 and 0% at Tn4.

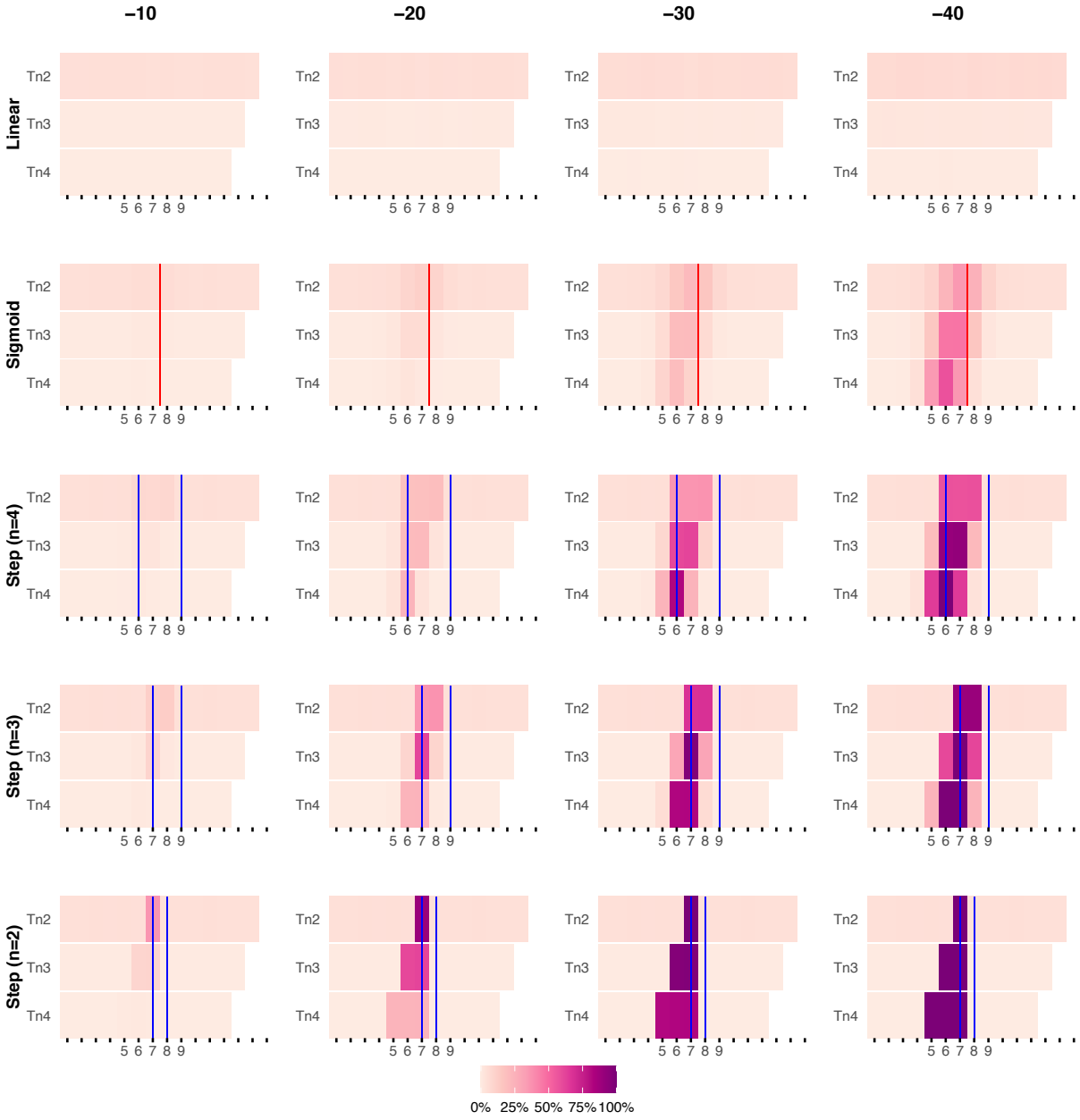
Sigmoid. Looking at the sigmoid change curves, where the exact starting point of change is less apparent, the DaRCI method started to pick up changes from an overall change of –30: about 20% for all three different duration thresholds. When the overall amount of change was –40, the DaRCI thresholds identified the correct period of increased discontinuity in about 35–58% of the simulations.

Step (n=4). For the third row of subplots in Figure 2, changes started to be noticeably detected in the –20 overall change model. Because the change was modeled to take place over four time points, the Tn4 threshold yields the highest specificity: 24% of changes were determined to start at time point 6, as modeled. At higher levels of overall change, this increased to 83% and 100%. This while in the overall change –20 models, changes over two (Tn2) and three points (Tn3) were still similar in accuracy: about 19% and 21% of simulations identified a change in the correct period, respectively. For Tn2, this improved to 36% at –30, and 58% at –40 overall change, while Tn3 accurately captured 62%, and 91% at those levels of overall change.

Step (n=3). When change was modeled as a step function over three time points, the DaRCI over three change points picked this up accurately at 62% at –20 points overall change, and 98% and 100% at larger changes (–30, –40). To compare, the DaRCI at Tn2 also showed about 37% accuracy at identifying the location of the mean shift at –20, but the identified changes were

Figure 2

Heatmap of DaRCI-detected starting points (t_{start}) of reliable changes in the simulated time series



Note. Visualization of the frequency of indicated starting points of reliable change across the 1000 simulations. Varying levels of overall change (–10 to –40) are shown on the general X-axis (see headers), and increasing discontinuity on the general Y-axis (from no shift within the time series, to an abrupt shift between two points). The red vertical lines indicate the midpoint of the sigmoid function, and the blue lines indicate the start and end of the step functions. Tn2 = (Da)RCI calculated for change between two points; Tn3 = DaRCI calculated for change over 3 points; Tn4 = DaRCI calculated for change over 4 points. Note that the location of t_{start} contributes to the value count, and the accuracy of the DaRCI must be inferred from whether the simulated change period falls within the range of the threshold's duration (Tn2, Tn3 and Tn4).

divided over time point 7 and 8, as the one-week duration could not capture the entire change period that was modeled. Conversely, for Tn4, the changes were harder to identify because often the overall required change (Tn4 = 25) to meet that threshold was not met (only in about 23% of cases at –20 decline). These durations improved when the overall change increased to –30: with about 68% of simulations indicating changes over two time points, and 82% finding shifts over four points as well.

Step (n=2). In the bottom row of subplots, where a mean shift was modeled as occurring between two time points, the RCI was still able to pick up the correct location of the shift for about 36% of cases at Tn2 and for about 9% at Tn3 at the lowest level of overall score reduction (–10). For the –20 change model, the DaRCI over two points picked up the change point accurately in 90% of simulations (compared to 62% at Tn3, and 24% at Tn4). At –30, only the Tn4 model sometimes did not pick up the modeled shift (the other thresholds picked up ~100%), yet in 82% of simulations it indicated a shift started somewhere in the range of time point 5 to 7. Finally, at the highest level of overall change (–40), the change was correctly identified in 100% of cases for all DaRCI thresholds.

Discussion

This paper provided a first illustration of a newly proposed method to identify reliable changes in symptoms of varying duration. Based on the current simulation study of symptom time series, it appears that the DaRCI is a valuable extension of the standard RCI for detecting reliable shifts over multiple time points in the course of treatment. The simulations showed that the DaRCI thresholds over more than two observations (Tn3 and Tn4) were able to pick up the modeled points and periods of discontinuity with high accuracy, especially when the overall change was large (30 points or more). The longer durations also showed fewer false positive values than the DaRCI for change over two points (Tn2, which is equivalent to the standard RCI, but set to a chosen time interval). The DaRCI over three time points (Tn3) seemed to be particularly well-suited to identifying both the relatively abrupt changes modeled with the step function over two and three observations, and the more

gradual changes in the sigmoid function and the slowest step function over four points. The DaRCI over four observations (Tn4) also performed well, even though it showed slightly more dispersion of the identified start points. This was not necessarily a problem of its performance, instead it indicates that the midpoint of a transition could be captured accurately, but the starting point of a change was less precisely estimated due to the larger range of the DaRCI at that duration. Similarly, the threshold for change over two points sometimes picked up reliable changes within the context of a larger overall shift (thus identifying a partial transition). To tease apart overlapping reliable changes, the period that shows relatively (to the time it took) the largest change, can be identified with the DaRCI by looking at which threshold an identified transition has the highest Z-score. Taken together, the results of this paper show that the DaRCI has the potential to explore symptom changes of different duration in the context of repeated outcome monitoring.

Apart from the accuracy of the DaRCI thresholds, the simulations also show that investigating reliable changes over different time periods with a purposely adapted index has the potential to uncover clinically relevant symptom changes that may be modest, step-by-step, but large overall. Employing only the standard RCI cut-off and testing difference scores without regard for time (any change that meets the criterion counts), or with a mere repetition of the same criterion for each increment (test the differences for point 1-2, point 2-3, etc., change is found only when those adjacent points meet the cut-off), would overlook these kinds of continued changes. Clinically, it can already be of interest to study whether multiple change thresholds are met by one individual, or which kinds of changes occur within a given study population (e.g., response patterns in depression [24,38,51], or mood shifts in bipolar disorder [52,53]). The DaRCI may be particularly useful for identifying a discontinuous period within the course of a larger change process. However, whether the DaRCI will be able to detect such periods of relevant change relies on the chosen time period between observations, and exploring the time scale of clinical change processes is still an important topic of study in and of itself [54–57]. However, given the extant literature on the importance of particular change patterns for the outcome of treatment (e.g., Refs. [10,18,25,34,39,40]), the DaRCI method

may provide a novel way to explore reliable changes of varying durations within the context of treatment.

One may note that the DaRCI thresholds becoming more demanding over longer periods, is statistically unexpected. Commonly, when more data is available, the additional power allows smaller changes to be detected (e.g., in a linear regression). While this is a fair observation, the aim is not to determine a statistically significant change [1,4,7]. Instead, the DaRCI may provide a simple method to add to a clinician's toolbox, which allows periods of reliable change to be uncovered within the course of a longer time series, even when little data is available. The DaRCI uses between-persons information (the SE_{m} from the instrument's norm group) to allow within-individual changes to be differentiated from measurement error. Certainly, some individuals may still show variations far beyond the expected range, seemingly showing reliable changes almost every other step; just like some individuals are more likely to vary very little over time, and never show changes that meet the threshold, even if they experience them as meaningful. Once more data is available, more refined methods may be useful in determining significant individual changepoints or overall change [44,46,58].

This method is not without some limitations. First is the challenge of setting a baseline duration for the RCI between two time points, from which the DaRCI calculated the extended thresholds. Researchers must decide for their study what the relevant periods of change may be, which may require pilot studies and in-depth clinical knowledge of the population and change process under study to come to an educated best guess. Another limitation is that although the DaRCI is very flexible and can be adapted to any chosen symptom measurement instrument, the RCI relies on clinical norm scores for optimal performance [4]. In our simulations we used the SCL-90 norms as a reference as they are based on a large sample (N = 5,621). Other instruments may not have such norm scores available, which makes calculations of the standard error of measurement and consequently the (Da)RCI potentially less accurate [1,7]. It is also worth remarking upon time period underpinning the $\boldsymbol{SE_{\mathrm{m}}}$: the reliability coefficient on which the $\boldsymbol{SE_{\mathrm{m}}}$ is based, is drawn from test-retest reliability (although internal consistency can be used as well), which would likely be based on a repeated assessment after several months [7]. This is quite different from the time periods (weeks) discussed in this paper

for the DaRCI. However, this problem may be minor, as the chosen $\boldsymbol{SE_m}$ is likely to have resulted in more conservative thresholds, as test-retest reliability would be higher if measurements occurred close together in time (and the $\boldsymbol{SE_m}$ smaller).

A strength of this method is that it has the potential for easy application in the clinical context. Once the thresholds have been calculated, clinicians can check whether new symptom assessments (e.g., as gathered with routine outcome monitoring) meet the shorter or longer duration thresholds of the DaRCI. Thus, using this method to examine within-person symptom time series for the presence changes of different durations, particularly while they are in treatment, could yield a novel view of change processes during therapy

Future research should aim to validate this method with clinical data, to provide insight into the kind of symptom changes that can be detected in real-world symptom assessments. It would further be of interest to expand the application of the DaRCI to other commonly used measurement instruments to test if it remains equally effective. Furthermore, comparing it to existing models of change within the course of treatment, such as "sudden gains" would be worthwhile, although the two methods may have slightly different objectives (i.e., identifying changes of various durations, versus change between therapy sessions).

To conclude, the DaRCI provides a simple adaptation of a well-established method for identifying reliable change [3,7,41], and may encourage researchers to consider exploring (discontinuous) symptom shifts of varying durations in the context of psychological treatment.

Acknowledgements

The author would like to thank Arnout Smit, Merijn Mestdagh and Francis Tuerlinckx for their critical input and help with the conception of the simulation study in this paper, and Evelien Snippe, Laura Bringmann and Tineke Oldehinkel for their encouragement and insightful text revisions.

References

1. Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*(3), 300–307. https://doi.org/10.1037/0022-006x.67.3.300

2. Ilardi, S. S., & Craighead, W. E. (1994). The role of nonspecific factors in cognitive-behavior therapy for depression. *Clinical Psychology: Science and Practice*, *1*(2), 138–155. https://doi.org/10.1111/j.1468-2850.1994.tb00016.x

3. Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

4. Jacobson, N. S., & Truax, P. (1991). Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. In *Journal of Consulting and Clinical Psychology* (Vol. 59, Issue 1). https://doi.org/10.1037/0022-006X.59.1.12

5. Maassen, G. H. (2000). Kelley's formula as a basis for the assessment of reliable change. *Psychometrika*, *65*(2), 187–197. https://doi.org/10.1007/BF02294373

6. Maassen, G. H., Bossema, E., & Brand, N. (2009). Reliable change and practice effects: Outcomes of various indices compared. *Journal of Clinical and Experimental Neuropsychology*, *31*(3), 339–352. https://doi.org/10.1080/13803390802169059

7. Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*(1), 60–70. https://doi.org/10.1207/s15327752jpa8201_11

8. Keller, M. B. (2003). Past, present, and future directions for defining optimal treatment outcome in depression. *JAMA*, *289*(23), 3152–3160. https://doi.org/10.1001/jama.289.23.3152

9. Tang, T. Z., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *67*(6), 894–904. https://doi.org/10.1037/0022-006X.67.6.894

10. Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C., Mocanu, S., Vocks, S., Schulte, D., & Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, *23*(1), 14–24. https://doi.org/10.1080/10503307.2012.693837

11. Tang, T. Z., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *67*(6), 894–904. https://doi.org/10.1037/0022-006X.67.6.894

12. Tang, T. Z., Luborsky, L., & Andrusyna, T. (2002). Sudden gains in recovering from depression: Are they also found in psychotherapies other than cognitive-behavioral therapy? *Journal of Consulting and Clinical Psychology*, *70*(2), 444–447. https://doi.org/10.1037/0022-006X.70.2.444

13. Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(1), 93–101. https://doi.org/10.1037/a0026455

14. Shalom, J. G., & Aderka, I. M. (2020). A meta-analysis of sudden gains in psychotherapy: Outcome and moderators. *Clinical Psychology Review*, *76*, 101827. https://doi.org/10.1016/j.cpr.2020.101827

15. Ilardi, S. S., & Craighead, W. E. (1994). The Role of Nonspecific Factors in Cognitive-Behavior Therapy for Depression. *Clinical Psychology: Science and Practice*, *1*(2), 138–155. https://doi.org/10.1111/j.1468-2850.1994.tb00016.x

16. Haas, E., Hill, R. D., Lambert, M. J., & Morrell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology*, *58*(9), 1157–1172. https://doi.org/10.1002/jclp.10044

17. Paul, R., Andlauer, T. F. M., Czamara, D., Hoehn, D., Lucae, S., Pütz, B., Lewis, C. M., Uher, R., Müller-Myhsok, B., Ising, M., & Sämann, P. G. (2019). Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical

prediction models. *Translational Psychiatry*, *9*(1), 1–15. https://doi.org/10.1038/s41398-019-0524-4

18. Stulz, N., Lutz, W., Leach, C., Lucock, M., & Barkham, M. (2007). Shapes of early change in psychotherapy under routine outpatient conditions. *Journal of Consulting and Clinical Psychology*, *75*(6), 864–674. https://doi.org/10.1037/0022-006X.75.6.864
19. aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review*, *32*(6), 510–523. https://doi.org/10.1016/j.cpr.2012.05.007
20. Barkham, M., Stiles, W. B., & Shapiro, D. A. (1993). The shape of change in psychotherapy: Longitudinal assessment of personal problems. *Journal of Consulting and Clinical Psychology*, *61*(4), 667–677. https://doi.org/10.1037/0022-006X.61.4.667
21. Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., & Harrington, H. (2020). Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the dunedin birth cohort study. *JAMA Network Open*, *3*(4), e203221. https://doi.org/10.1001/jamanetworkopen.2020.3221
22. Ebner-Priemer, U. W., Kubiak, T., & Pawlik, K. (2009). Ambulatory Assessment. *European Psychologist*, *14*(2), 95–97. https://doi.org/10.1027/1016-9040.14.2.95
23. Schiepek, G., Aichhorn, W., Gruber, M., Strunk, G., Bachler, E., & Aas, B. (2016). Real-time monitoring of psychotherapeutic processes: Concept and compliance. *Frontiers in Psychology*, *7*, 604. https://doi.org/10.3389/fpsyg.2016.00604
24. Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2013). Nomothetic and idiographic symptom change trajectories in acute-phase cognitive therapy for recurrent depression. *Journal of Consulting and Clinical Psychology*, *81*(4), 615–626. https://doi.org/10.1037/a0032879
25. Hayes, A. M., Laurenceau, J.-P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in

psychotherapy. *Clinical Psychology Review*, *27*(6), 715–723. https://doi.org/10.1016/j.cpr.2007.01.008

26. Gelo, O. C. G., & Salvatore, S. (2016). A dynamic systems approach to psychotherapy: A meta-theoretical framework for explaining psychotherapy change processes. *Journal of Counseling Psychology*, *63*(4), 379–395. https://doi.org/10.1037/cou0000150
27. Schiepek, G. (2009). Complexity and nonlinear dynamics in psychotherapy. *European Review*, *17*(02), 331. https://doi.org/10.1017/S1062798709000763
28. Helmich, M. A., Wichers, M., Olthof, M., Strunk, G., Aas, B., Aichhorn, W., Schiepek, G., & Snippe, E. (2020). Sudden gains in day-to-day change: Revealing nonlinear patterns of individual improvement in depression. *Journal of Consulting and Clinical Psychology*, *88*(2), 119–127. https://doi.org/10.1037/ccp0000469
29. Tang, T. Z., Luborsky, L., & Andrusyna, T. (2002). Sudden gains in recovering from depression: Are they also found in psychotherapies other than cognitive-behavioral therapy? *Journal of Consulting and Clinical Psychology*, *70*(2), 444–447. https://doi.org/10.1037/0022-006X.70.2.444
30. Aderka, I. M., & Shalom, J. G. (2021). A revised theory of sudden gains in psychological treatments. *Behaviour Research and Therapy*, *139*, 103830. https://doi.org/10.1016/j.brat.2021.103830
31. Tadić, A., Helmreich, I., Mergl, R., Hautzinger, M., Kohnen, R., Henkel, V., & Hegerl, U. (2010). Early improvement is a predictor of treatment outcome in patients with mild major, minor or subsyndromal depression. *Journal of Affective Disorders*, *120*(1–3), 86–93. https://doi.org/10.1016/J.JAD.2009.04.014
32. Lutz, W., Stulz, N., & Köck, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of Affective Disorders*, *118*(1–3), 60–68. https://doi.org/10.1016/J.JAD.2009.01.019

33. Abel, A., Hayes, A. M., Henley, W., & Kuyken, W. (2016). Sudden gains in cognitive–behavior therapy for treatment-resistant depression: Processes of change. *Journal of Consulting and Clinical Psychology*, *84*(8), 726–736. https://doi.org/10.1037/ccp0000101

34. Thompson, M., Thompson, L., Gallagher-Thompson, D., & Alto, P. (1995). Linear and nonlinear changes in mood between psychotherapy sessions: Implications for treatment outcome and relapse risk. *Psychotherapy Research*, *5*(4), 327–336. https://doi.org/10.1080/10503309512331331436

35. Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology & Psychotherapy*, *8*(4), 231–242. https://doi.org/10.1002/cpp.286

36. Lambert, M. J. (2005). Early response in psychotherapy: Further evidence for the importance of common factors rather than “placebo effects”. *Journal of Clinical Psychology*, *61*(7), 855–869. https://doi.org/10.1002/jclp.20130

37. Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy Clin. Psychol. Psychother*, *9*, 149–164. https://doi.org/10.1002/cpp.333

38. Rubel, J., Lutz, W., Kopta, S. M., Köck, K., Minami, T., Zimmermann, D., & Saunders, S. M. (2015). Defining early positive response to psychotherapy: An empirical comparison between clinically significant change criteria and growth mixture modeling. *Psychological Assessment*, *27*(2), 478–488. https://doi.org/10.1037/pas0000060

39. Schiepek, G., Aichhorn, W., & Schöller, H. J. (2017). Monitoring change dynamics: A nonlinear approach to psychotherapy feedback. *Chaos and Complexity Letters*, *11*(3), 355–375.

40. Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2016). Defined symptom-change trajectories during acute-phase cognitive therapy for depression predict better longitudinal

outcomes. *Behaviour Research and Therapy*, *87*, 48–57. https://doi.org/10.1016/j.brat.2016.08.008

41. Ferrer, R., & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, *26*(2), 370–383. https://doi.org/10.1037/a0035419
42. Slofstra, C., Nauta, M. H., Bringmann, L. F., Klein, N. S., Albers, C. J., Batalas, N., Wichers, M., & Bockting, C. L. H. (2018). Individual negative affective trajectories can be detected during different depressive relapse prevention strategies. *Psychotherapy and Psychosomatics*, *87*(4), 243–245. https://doi.org/10.1159/000489044
43. Maric, M., de Haan, E., Hogendoorn, S. M., Wolters, L. H., & Huizenga, H. M. (2015). Evaluating statistical and clinical significance of intervention effects in single-case experimental designs: An spss method to analyze univariate data. *Behavior Therapy*, *46*(2), 230–241. https://doi.org/10.1016/j.beth.2014.09.005
44. Albers, C., & Bringmann, L. F. (2020). Inspecting gradual and abrupt changes in emotion dynamics with the time-varying change point autoregressive model. *European Journal of Psychological Assessment*, *36*(February), 492–499. https://doi.org/10.1027/1015-5759/a000589
45. Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., & Ceulemans, E. (2017). Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods. *Behavior Research Methods*, *49*(3), 988–1005. https://doi.org/10.3758/s13428-016-0754-9
46. de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*(2), 165–185. https://doi.org/10.1037/a0031037
47. Arrindell, W. A., & Ettema, J. H. M. (2003). *SCL-90: Manual to a multidimensional psychopathologyindicator [SCL-90: Handleiding bij een multidimensionele psychopathologie-indicator]* (Pearson, Ed.; 2nd ed.). Swets & Zeitlinger.
48. Derogatis, L. R. (1977). *SCL-90-R: administration, scoring and procedures manual-I for the R(evised) version*. John Hopkins University School Medicine.

49. Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*(3), 329–343. https://doi.org/10.1093/clipsy.9.3.329

50. Gloaguen, V., Cottraux, J., Cucherat, M., & Ivy-Marie Blackburn. (1998). A meta-analysis of the effects of cognitive therapy in depressed patients. *Journal of Affective Disorders*, *49*(1), 59–72. https://doi.org/10.1016/S0165-0327(97)00199-7

51. Korf, J. (2014). Delayed mood transitions in major depressive disorder. *Medical Hypotheses*, *82*(5), 581–588. https://doi.org/10.1016/j.mehy.2014.02.015

52. Bos, F. M., Schreuder, M. J., George, S. V., Doornbos, B., Bruggeman, R., van der Krieke, L., Haarman, B. C. M., Wichers, M., & Snippe, E. (2021). Anticipating manic and depressive shifts in patients with bipolar disorder using early warning signals. *Submitted*.

53. Kramlinger, K. G., & Post, R. M. (1996). Ultra-rapid and ultradian cycling in bipolar affective illness. *British Journal of Psychiatry*, *168*(3), 314–323. https://doi.org/10.1192/bjp.168.3.314

54. Strunk, G., & Lichtwarck-Aschoff, A. (2019). Therapeutic chaos. *Journal for Person-Oriented Research*, *5*(2), 81–100. https://doi.org/10.17505/jpor.2019.08

55. Hayes, S. C., Hofmann, S. G., Stanton, C. E., Carpenter, J. K., Sanford, B. T., Curtiss, J. E., & Ciarrochi, J. (2019). The role of the individual in the coming era of process-based therapy. *Behaviour Research and Therapy*, *117*, 40–53. https://doi.org/10.1016/j.brat.2018.10.005

56. Kazdin, A. E. (2001). Progression of therapy research and clinical application of treatment require better understanding of the change process. *Clinical Psychology: Science and Practice*, *73*(4), 143–151. https://doi.org/10.1093/clipsy/8.2.143

57. Mahoney, M. J. (2004). Human change processes and constructive psychotherapy. In *Cognition and psychotherapy, 2nd ed.* (pp. 5–24). Springer Publishing Co.

58. de Vries, R. M., Hartogs, B. M. A., & Morey, R. D. (2014). A tutorial on computing bayes factors for single-subject designs. *Behavior Therapy*, *46*(6), 809–823. https://doi.org/10.1016/j.beth.2014.09.013