

When Average Isn't Good Enough: Identifying Meaningful Subgroups in Clinical Data

Andrew T. Gloster¹, Matthias Nadler^{1,2}, Victoria Block^{1,3}, Elisa Haller¹, Julian Rubel⁴, Charles Benoy^{1,5}, Jeanette Villanueva^{1,6}, Klaus Bader⁷, Marc Walter⁸, Undine Lang⁷, Stefan G. Hofmann^{9,10}, Joseph Ciarrochi¹¹, Steven C. Hayes¹²

¹ Division of Clinical Psychology & Intervention Science, Department of Psychology, University of Basel, Basel, Switzerland

² Center for Innovative Finance, University of Basel, Basel, Switzerland

³ Psychiatric Hospital Sonnenhalde, Riehen, Switzerland

⁴ Department of Clinical Psychology and Psychotherapy, University of Giessen, Giessen, Germany

⁵ Rehaklinik, Centre Hospitalier Neuro-Psychiatrique Luxembourg (CHNP), Ettelbruck, Luxembourg

⁶ Psychiatric Center Wetzikon (Clenia Schlössli AG), Wetzikon, Switzerland

⁷ University Psychiatric Clinics (UPK), University of Basel, Basel, Switzerland

⁸ Psychiatric Services Aargau (PDAG), Windisch, Switzerland

⁹ Department of Psychology, Boston University, Boston, United States of America

¹⁰ Alexander von Humboldt Professor, Department of Clinical Psychology, Philipps-Universität Marburg, Marburg, Germany

¹¹ Institute for Positive Psychology and Education, Australian Catholic University, Sydney, Australia

¹² Department of Psychology, University of Nevada, Reno, United States of America

Author's Note:

Correspondence concerning this article should be addressed to Andrew T. Gloster, (ORCID: 0000-0002-3751-0878) University of Basel, Department of Psychology, Division of Clinical Psychology and Intervention Science, Missionsstrasse 62A, CH-4055 Basel (andrew.gloster@unibas.ch).

Abstract

Background: Clinical data are usually analyzed with the assumption that knowledge gathered from group averages applies to the individual. Doing so potentially obscures patients with meaningfully different trajectories of therapeutic change. Needed are “idionomic” methods that first examine idiographic patterns before nomothetic generalizations are made. The objective of this paper is to test whether such an idionomic method leads to different clinical conclusions.

Methods: 51 patients completed weekly process measures and symptom severity over a period of eight weeks. Change trajectories were analyzed using a nomothetic approach and an idiographic approach with bottom-up clustering of similar individuals. The outcome was patients’ well-being at post-treatment.

Results: Individuals differed in the extent that underlying processes were linked to symptoms. Average trend lines did not represent the intraindividual changes well. The idionomic approach readily identified subgroups of patients that differentially predicted distal outcomes (well-being).

Conclusions: Relying exclusively on average results may lead to an oversight of intraindividual pathways. Characterizing data first using idiographic approaches led to more refined conclusions, which is clinically useful, scientifically rigorous, and may help advance individualized psychotherapy approaches.

Keywords: intraindividual differences, idionomic analysis, nomothetic, processes

Introduction

Most studies in medical and behavioral science use statistics with the assumption that the conclusions derived from calculated mean-level results apply at least probabilistically to most members of the sample. Said in another way, it is assumed that measurement of central tendencies and variability between people can be used to model central tendencies and variability within persons.

Measurement theorists have long understood that this assumption could be wrong (Cattell, 1952) and it has sometimes been challenged in areas such as developmental psychology (Wohlwill, 1973), epigenetics (Gottlieb, 1992), or behavior analysis (Peters & Sidman, 1961). In the main, these intellectual objections have made little impact and most of the modern medical and behavioral science is based on normative statistical approaches.

Peter Molenaar (2004) appears to be the first to have linked this issue to a well-established concept in the physical sciences: ergodicity (Birkhoff, 1931; Boltzmann, 1885; von Neumann, 1932). Although its focus is on modeling the distribution of elements in space and time, in a more abstract sense the ergodic theorem describes the conditions under which data applying to a collective also applies to components of that collective. The theorem reveals that such an extension is only proper if the elements are ergodic. A simple way to think of ergodicity in a fashion that relates to medical and behavioral science is that for ergodic events the means and standard deviations of all individuals across time and for all cross-sectional samples of collections of individuals at any given point in time will be the same (Gates et al., in press). That is sometimes true and sometimes not in the physical sciences (Cherstvy et al., 2019; Galenko & Jou, 2019; Kuehn & Abrahamson, 2020), but it is rare to the point of being absent in the human and life sciences (Horst, 2008; Kelderman & Molenaar, 2007).

A small number of empirical studies have shown that nomothetic and idiographic analyses can lead to different conclusions (Fisher et al., 2018; Miller & van Horn, 2007; Molenaar & Campbell, 2009) but these ideas have not been widely adopted in intervention research. One reason may be that these studies have generally been based on complex methods or with data requirements that are far from typical research practices in the medical and behavioral sciences. For example, while complex network analyses, p factor analyses, replicated multivariate time series analyses, and other methods can be applied to longitudinal data sets when scores of within person observations are available, such data are usually not available from most randomized controlled trials, which typically have only a few measurement points. The net effect has been that theoretical alarms about the applicability of normative analyses have seemed empirically remote, and empirical tests of the need for idiographic data are both uncommon and readily set aside.

This issue has recently taken on special urgency in psychological intervention and the health sciences writ large because research attention is increasingly being focused on identifying processes of change that meaningfully and reliably apply to those presenting for treatment, in such areas as process-based therapy or personalized medicine. Focusing on processes of change creates a conundrum. On the one hand, a change process cannot be ergodic because stationarity is a key feature of ergodic events (Molenaar, 2013); on the other hand, knowledge drawn from a truly idiographic approach cut off from nomothetic conclusions cannot be applied to others. One possible solution would be for analyses to initially be based on idiographic conclusions, examined in the context of within person variability. These individual analyses could then be gathered into nomothetic generalizations that are ultimately shown to improve idiographic fit – what has been called an “idionomic” rather than normative approach (Hayes & Hofmann, 2021).

The absence of ergodicity does not invalidate nomothetic generalizations provided that generalizations are derived from intra-individual variation. It is important to examine ways to begin analysis entirely idiographically using commonly available data before gathering individuals into similar subpopulations and examining whether doing so improves the fit between individual behavior and subpopulation average behavior. Applied researchers may believe that they are doing something similar with moderation analysis, cluster analysis, growth curve modeling, longitudinal mixed regression models, and so on but these standard methods interpret individuals in the context of inter-individual variability first – precisely the step that may not be justified if the ergodicity assumption is violated (Molenaar, 2013). In contrast, if analyses that begin with intra-individual variability were routinely successful, intervention scientists might be able to select targeted interventions tailored to a more homogenous group or subgroup of similar patients. It remains empirically untested, however, whether such an approach would render more accurate and useful predictions in clinical trials.

Theoretically, subgroups can be based on any observable, treatment-relevant variable ranging from scores on a questionnaire, to genetic polymorphisms, to nurse observations. The growing evidence in support of a focus on processes of change, however, suggests that the treatment relevance of building subgroups will be increased to the degree that the grouping variable a) stipulates a hypothesized mechanism of action that can be assessed and manipulated, and b) predicts a clinically relevant outcome; e.g., visual acuity, white blood cell count, mortality, depression symptoms, well-being, and so on (Gloster & Karekla, 2020; Hofmann & Hayes, 2019).

The present study implemented this process-based strategy with a sample of psychiatric patients who were assessed on hypothesized mechanisms of action consistent with state-of-the-art theories of change during psychotherapy (Gloster et al., 2017; Hayes et al.,

2019; Kashdan & Rottenberg, 2010) in addition to the measurement of typical syndromal features. We included variables previously shown to be active mechanisms of change in psychotherapy using nomothetic methods, such as mindfulness, relating differently to one's thinking, and acting on one's values (Hayes et al., 2020).

Different patients will respond differently to exercises designed to change such processes: e.g., some will improve their mindfulness skills, others may not change at all or even worsen in their ability to be mindful. Because change occurs over time, it is necessary to collect longitudinal data to model change trajectories (Gloster et al., 2014) and to assess it in terms of intra-individual variability. In the long run, daily diary apps or wearable automated data collection may allow the routine use of longitudinal models that are based on a series of several dozen longitudinal data points (Gates et al., 2017; Gates & Molenaar, 2012), but at the moment, data of that kind is rarely part of existing research designs. In the present study, we examined the application of a relatively simple idiographic approach built upon a limited number of data points of the kind routinely collected in clinical practice. In this approach, each individual was characterized in a bottom-up way that involved estimating individual effects that were uninfluenced by other individual effects. These were then clustered in a fashion that never treated the individual as “error.”

In the present study, this idionomic strategy was used to examine three key aims: 1) test whether assumptions of ergodicity are met or violated in clinical data; 2) test whether building subgroups using a bottom-up idiographic procedure before nomothetic analysis leads to different conclusions about underlying change than traditional nomothetic approaches; and 3) test whether idiographically identified bottom-up subgroups predicted a clinically meaningful distal outcome (namely, well-being), suggesting possible advantages for a more idionomic approach. These questions were examined using longitudinal data from a real

clinical sample with time series data. Examining these questions under real world conditions was designed to maximize the chance that the results are clinically relevant.

Methods

Participants

All methods were approved by the local ethics committee (EKNZ-165/13) and all participants completed an informed consent. Data were derived from a transdiagnostic controlled effectiveness clinical trial focused on treating patients diagnosed with a DSM-IV disorder in an inpatient setting using Acceptance and Commitment Therapy (ACT), a common transdiagnostic intervention method (Ciarrochi et al., 2023; Sahdra et al., 2022). Participants for this study were a subgroup of in-patients who completed all weekly assessments during active treatment from week 1 to week 8 on symptoms (proximal outcome measure) and a process measure of psychological flexibility (Psy-Flex; Gloster et al., 2021) without gaps or missing items (n=51). Additionally, as treatment usually lasted 8-12 weeks, two patients who stayed in treatment longer than 14 weeks were removed as outliers for these analyses. Table 1 presents detailed sample statistics.

Table 1

Sample Characteristics

	Total N=51
Age	
Mean (SD)	33.61 (10.14)
Sex	
Females N (%)	22 (41.51)
Diagnoses N (%)	
Mood Disorder	30 (58.82)
Anxiety Disorder	39 (76.47)

Obsessive Compulsive Disorder	20 (39.22)
Other Disorder	20 (39.22)
Two or more diagnoses	36 (70.59)

Assessments

Assessments included six process variables, one proximal outcome variable and one distal outcome variable. The process variables were completed weekly using the Psy-Flex (Gloster et al., 2021). The Psy-Flex measures skills of psychological flexibility and is sensitive to change produced by the intervention used in the original study. The items query about: 1) being present (“Even if I am somewhere else with my thoughts, I can focus on what's going on in important moments”), 2) own values (“I engage thoroughly in things that are important, useful, or meaningful to me”), 3) being engaged (“I determine what's important for me and decide what I want to use my energy for”), 4) steady self (“Even if thoughts and experiences are confusing me I can notice something like a steady core inside of me”), 5) defusion (“I can look at hindering thoughts from a distance without letting them control me”), and 6) acceptance (“If need be, I can let unpleasant thoughts and experiences happen without having to get rid of them immediately”). Weekly reports of the level of suffering caused by symptoms as assessed with one self-constructed item served as the proximal outcome variable (“During the last seven days, how much did you suffer because of your symptoms?”). When completing weekly assessments, participants reported on their experiences for the previous seven days. The distal outcome variable measured patients’ subjective level of well-being as assessed by the Mental Health Continuum – Short Form (MHC-SF) and was collected pretreatment and post-treatment (Keyes et al., 2008).

Nomothetic Analytical Approach

To test the nomothetic approach, we calculated the aggregated interindividual measurements for each variable at the group level (i.e., grouping all participants together for statistical analysis). The following statistics were calculated: Item mean change and standard deviations were calculated for each individual and then aggregated 1) per item per time point (i.e., week); 2) weekly mean change per item; and 3) baseline to posttreatment change by calculating the difference of week 1 and week 8 per item. A linear model was created for each item estimating a linear trend (slope and R-squared of the linear model). For better interpretation, the linear models for each item were visualized.

Test of the Ergodic Assumption

Ergodicity was formally examined based on the extent to which relationships between the processes and symptoms significantly varied within a person. Multi-level analyses were calculated using the lme4 package (Bates et al., 2015), nesting all observations within person. For every link between process and outcome, we compared a multi-level model that assumed random intercepts (allowing people to differ on the dependent variable) with one that assumed both random intercepts and slopes. A significant difference indicates that slopes between process and outcome differed between people, or that the within-person relationship between process and outcome differed from person to person.

Idiomatic Analytical Approach

In order to examine change using an idiomatic strategy, we first examined each individual's change trajectory over the eight weeks of observations. These individual change trajectories were then used to determine whether there were groups of individuals that had

similar trajectories using cluster analysis to arrive at these nomothetic sub-populations. No averages across individuals or traditional estimates of inter-individual variation were built into the clustering processes used.

To conduct this analysis, we first calculated trend/ regression lines for each individual and item separately. Therefore, no error term was calculated at the group level. Intra-individual trends are a discrete aspect of intra-individual variation, but it should be noted that intra-individual error terms can be also calculated for each individual. For example, intra-individual error terms can assess the degree to which the rate of change (slope) was not equidistant between time points for that participant. In the present study, such error terms factored into the analytic steps only in our assessment of ergodicity (e.g., Figure 4 below).

Because we were ultimately interested in identifying meaningful clinical groups, we considered both an individual's baseline value and their slope of change in subsequent steps. In order to facilitate weighting and interpretation, raw values (i.e., 1-5) were transformed so that values would be between 0 and 1 for the intercept and slope. This transformation did not alter the actual distance between individual's trend lines and no error term was used. To achieve this, we construct a normalized weighted distance based on the intercept and slope of the linear model for each participants' time series:

$$d(x, y) = |\alpha * (norm(x_{intercept}) - norm(y_{intercept}))| + \\ |\beta * (norm(x_{slope}) - norm(y_{slope}))|$$

The weights for the intercept and slope are α and β respectively and normalization transforms the values for intercept and slopes to values between 0 and 1 across all samples (i.e., the highest intercept is 1 and the lowest intercept is 0).

Choosing a Distance Measure

Clustering observations (i.e., individuals) requires a distance measure to determine the similarity of observations. Our observations are short time series and there are different ways

to calculate the similarity of two time series. In the absence of a universal method to compare time series and because the choice of distance measure has an impact on the result of the analysis, it is important to be very clear about the properties of the observations that were considered.

For our analysis, we needed a distance measure that captured two key metrics: The initial magnitude of an item (severity level) and the trend the item follows over the eight weeks (change). The latter was given more importance than the first a priori in this manuscript, in part because the intercept influences the trajectory in the form of ceiling/ basement effects and regression to the norm (i.e., a high intercept leaves little or no room for gains, low intercept increases the likelihood of gains).

Choosing Parameters for the Clustering

When choosing a distance measure, one can think of each individual's multiple observations over time as a single observation, having a location based on its regression parameters. The distance between two such observations is then given as the distance between their respective locations. To calculate the distance between two clusters, all the observations in both clusters have to be considered according to a linkage criterion. Our goal was to choose the linkage criterion resulting in the most robust and well-formed clustering structure given our data and our clustering algorithm. From the commonly used linkage criteria - single, complete, average, centroid, and Ward's (Jarman, 2020) - that we tested, Ward's minimum variance method (Ward, 1963) consistently resulted in the highest agglomerative coefficient (AC) by minimizing the total within-cluster variance. For our chosen clustering method, Agglomerative Nesting (AGNES) clustering, the clustering structure can be measured with the AC. The AC reveals the mean of the normalized lengths at which clusters are formed, and higher values for the AC indicate a stronger clustering structure (Gates et al., 2017).

When choosing the parameters to measure distance measures, weights to the α - and β -coefficient must be assigned. As we wanted to compare linear regression lines with each other, we needed to assign the parameters for intercept and slope, respectively. For reasons explained above, we assigned the slope more weight than the intercept, so that linear regressions with similar slopes were more likely to be grouped together as long as their intercepts are not too different. This is a practical way of being able to compare individual trajectories.

For the parameters α and β , we found that the values $\alpha = 1$ and $\beta = 3$ separated our time series in a meaningful way. These numbers were obtained by testing different values for α and β and comparing the resultant figures via visual inspection for which slope and intercept had the largest impact on the categorization (see supplementary material section A for examples). When participants showed similar changes over time, they were more likely to be grouped in a cluster by the function. When defining the parameters for distance, we used clustering methods that employed the Manhattan distance measure (Gonzalez & Palais, 1962), which measured the actual distance traveled between points (around the corners), and not the Euclidian distance (direct).

Choosing a Clustering Approach

In order to determine groups of individuals with similar change patterns we used clustering procedures. Clustering procedures must satisfy at least the following properties described by Xu and Tian (2015):

- (1) Instances in the same cluster must be as similar as possible
- (2) Instances in different clusters must be as different as possible
- (3) Measurements for similarity and dissimilarity must be clear and pragmatic

Based on these axioms, there are different methods that can be used to make clustering decisions. Approaches that start with each observation as its own cluster are commonly referred to as Hierarchical Clustering. To choose a metric, we ran our data through

the following clustering sequences: AGNES and Divisive Clustering (DIANA). The outcome of the two approaches were compared to choose the more appropriate method, as literature has not yet defined which is preferable in all instances (Edelbrock & McLaughlin, 1980). DIANA is a top-down method of hierarchical clustering that uses a reversed algorithm, starting with the assumption that all data points belong to the same cluster. In each iteration, the largest available cluster is then split into two clusters until each observation represents one cluster. The algorithm therefore needs $n-1$ iterations to complete. How clusters are split depends on how the parameters of dissimilarity are defined. AGNES, conversely, is a bottom-up approach commonly used to find and group similar observations in a set of data. It creates clusters in an iterative approach, starting out with as many clusters as there are observations, and with each step combining those that are most similar to each other to form a bigger cluster. In each iteration, the most similar clusters are combined until in the last step, all observations are merged into one cluster. Generally, AGNES yields more balanced clusters than DIANA. So far, there is no established way of deciding the best algorithm, which is why one has to define and justify the chosen parameters carefully. The results are then evaluated by comparing the cluster sizes on the one hand and visually checking the groups' trajectories on the other hand. Comparing the DIANA and AGNES clustering algorithms showed that the AGNES algorithm yielded more balanced clusters for our data (for an example of DIANA clustering, see Supplements B and C). Hence, we used AGNES for the idiographic clustering. To interpret the result of the clustering visually, Dendrograms, graphs that show the iteration like the root-system of a tree, were inspected. It is important to note that depending on the data and the clustering parameters, the most fitting clustering algorithm may differ.

Choosing the Number of Clusters

A four-cluster solution was chosen on the assumption that there are four possible therapy outcomes: slight and strong increase of parameters, and slight and strong decrease of

parameters. It would also be possible to base the number of clusters on a statistic measure like the silhouette plot, but there are no objective criteria for the correct number of clusters to choose.

Prediction of Distal Outcome

In order to examine whether the idiomically derived clusters of patients predicted clinically useful information nomothetically, we subsequently tested whether the clusters differed with respect to a distal outcome. Specifically, we tested whether patients in the clusters subsequently resulted in meaningfully different values on well-being. This was implemented by adding each patient to one of four factor levels based on their AGNES-derived groups for each process variable and then testing differences using an omnibus ANOVA. As the sample sizes of the resulting clusters were underpowered for such analyses, we concentrated on effect sizes using Cohen's f , an effect size measure used for ANOVA using pooled standard deviations.

Results

Nomothetic Approach

Results (see Table 2) showed that, on average, the *seven observed items* either revealed slopes indicative of some improvement over time, or no change. The linear models per item showed varying slopes, with some regression lines' slope being close to zero (e.g., item 3 - Being engaged; item 1 - Being present), suggesting that on average, no change had happened over the eight weeks that were measured. The weekly mean difference per item was limited for all items, with the highest score accounting for a 2% change over one week on the Psy-Flex scale (from 1-5). R-squared values per item ranged from .02 to .74. As r-squared values are lower, the poor fit of a purely linear nomothetic model indicates that there may be patterns in the data that are being overlooked due to the aggregating nature of nomothetic analysis.

Table 2

Nomothetic changes and results of linear models (N=51)

	weekly mean difference	Pre-post mean	Standard deviation	rmssd	slope	r-squared
1. Being present	-0.006	-0.039	0.84	0.84	-0.006	0.020
2. Own values	0.076	0.529	0.83	0.87	0.049	0.505
3. Being engaged	0.017	0.118	0.81	0.75	0.009	0.229
4. Steady self	0.084	0.588	1.07	0.96	0.066	0.738
5. Defusion	0.090	0.627	0.96	0.98	0.057	0.525
6. Acceptance	0.081	0.569	0.93	0.94	0.057	0.712
7. Symptoms	-0.036	-0.036	1.10	0.95	-0.022	0.480

Note. Weekly mean differences calculated across participants and between weeks. Pre-post mean calculated by subtracting overall mean value of week 8 from overall mean value of week 1. Standard deviation (SD) reflects inter-individual SD across all weeks. Rmssd, slope, and r-squared calculated by setting up a linear model.

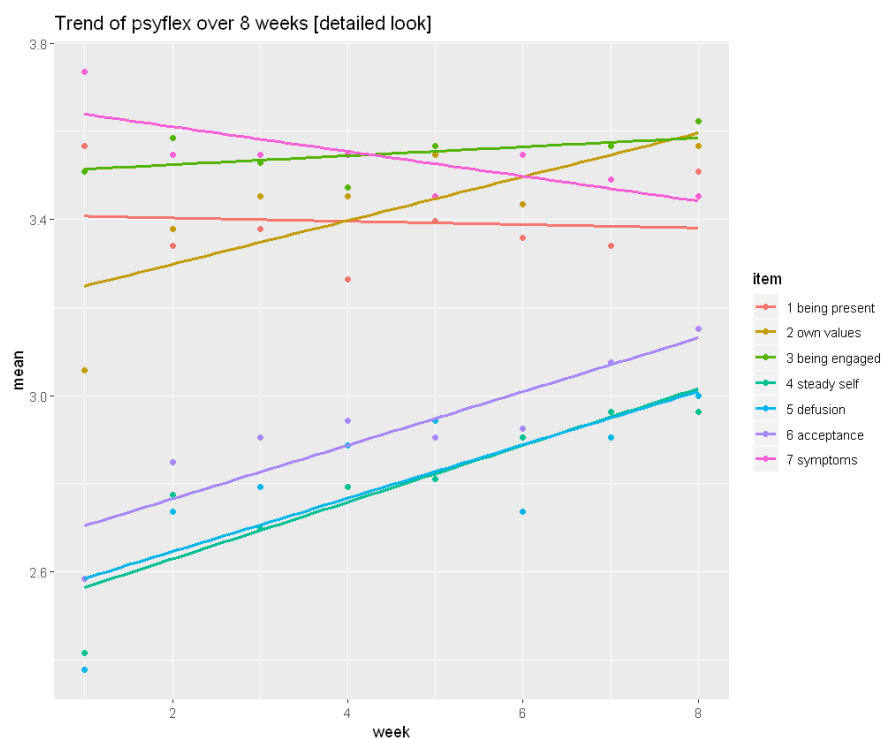


Fig. 1 Average trend lines for items measuring psychological flexibility and symptoms

Figure 1 shows the weekly means and the linear model for the items plotted over eight weeks. On average, “symptoms” shows a negative trend, the processes of “being present” and “being engaged” show no significant trends, and the remaining process items (“own values”, “acceptance”, “defusion”, and “steady self”) show positive trends.

Often, time-series data is analyzed using multi-level models, because such data tend to contain both missing data and an unequal number of observations (Snijders & Bosker, 2012), which is not the case here. Multi-level models consider inter-individual variance as a metric for group level estimates, however they make the assumption of normally distributed individual trajectories around the average intercept and slope, resulting in a restrictive model. To exemplify this, consider the first item, “being present”. As seen in the left side of Figure 2, the mean group level change for this item showed that minimal change across time occurred. The right side of Figure 2 shows an individual regression line for each participant in the sample (trends instead of point measures are displayed for better readability of the plot). A very different picture emerges when all 51 of the individual trend lines are contrasted with the aggregated trend line.

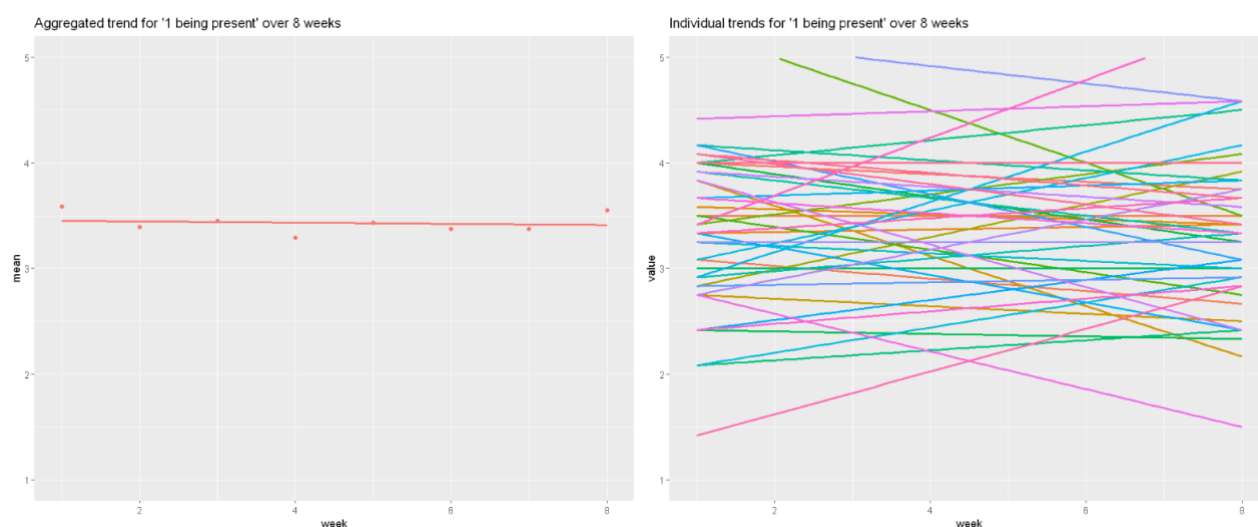


Fig. 2 Individual regression lines vs. aggregated regression line for the item “Being present”

The nomothetic measurement not only fails to describe the individual trends, but it is actively misleading in that it suggests that there is very little change over time. For many individuals that is hardly the case. The observed inter-individual standard deviation (see Table 2) hints that there is more going on, but standard deviation alone does not consider that each trend is a time series, nor the degree of intra-individual variability.

In order to examine the generalizability of this pattern, Figure 3 repeats this analysis with the item measuring symptoms. The same pattern is seen. Indeed, if we were to continue item by item, it is universally true that the nomothetic pattern fails to characterize most of the sample. For the aggregated and individual trend lines for the remaining items, please see supplementary section D.

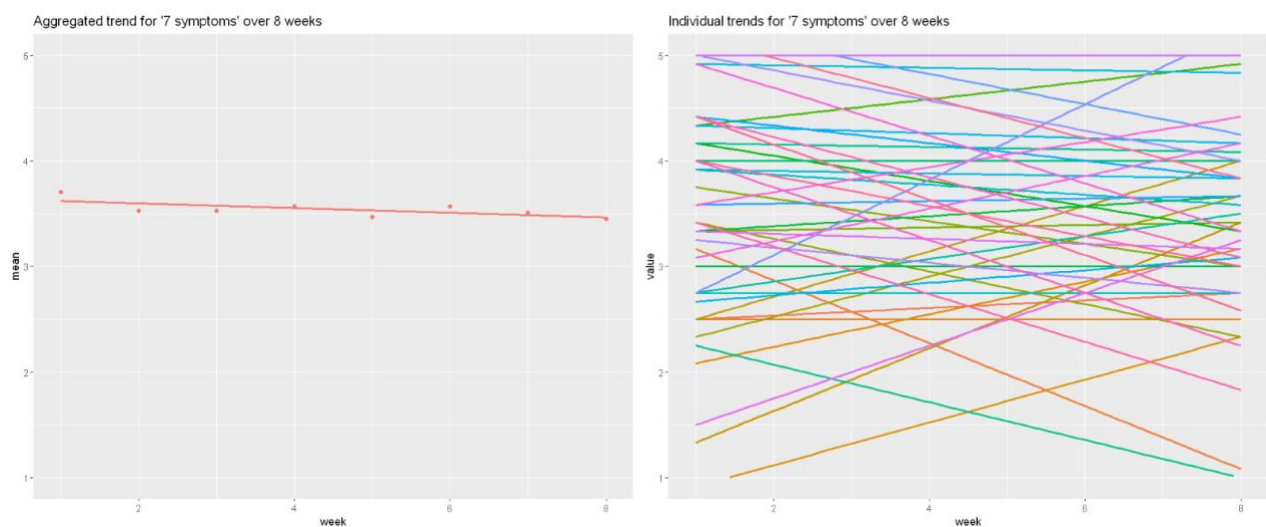


Fig. 1 Individual regression lines vs. aggregated regression line for the item "Symptoms"

Test of Ergodic Assumption

If we assume ergodicity, then the link between process and symptoms should not significantly differ between people. We evaluated this possibility using multi-level models that estimated intercepts and betas (relationships between process and symptoms) for each person. In Table 3, we present the fixed effect (average), and the 10th and 90th percentile of

the distribution of participant betas. The negative fixed effects indicate that all weekly processes were generally associated with lower symptoms, though the fixed effect for valuing was the only process to reach traditional statistical significance. However, the Chi-square tests of varying slopes were all significant, indicating that the relationship between the weekly process and weekly symptoms significantly varied from person to person. This within person variance appeared to be largest for *Steady self* and *Defusing*. Although processes were generally beneficial (negative fixed effects), there was a subset of people for whom the processes were either inert or possibly associated with worse symptoms, especially for Steady Self (Upper beta = .28) and Defusing (upper beta =.24).

Table 3

Betas and range of within person associations between processes and negative symptoms

Process	%10	Fixed Effect	%90	Chisq
Accepting	-0.33	-0.12	0.14	10.03**
Engaging	-0.24	-0.09	0.14	14.22***
Valuing	-0.28	-0.13*	0.11	16.44***
Being Present	-0.32	-0.07	0.07	15.08***
Steady Self	-0.37	-0.12	0.28	25.65***
Defusing	-0.44	-0.12	0.24	38.5***

Note: * $p < .05$; ** $p < .01$; *** $p < .001$; Chi-square tests that the relationship between process and symptoms varied by participant.

As an example of the implications of a lack of ergodicity, the within-person relationship between defusing and symptoms is illustrated in Figure 4. We can see that for some people, higher defusing is significantly associated with worse symptoms (e.g., see

participant 6 and 22, for whom intra-individual standard error bars do not overlap with zero).

For others (e.g., participant 42 and 44), defusing is associated with fewer symptoms.

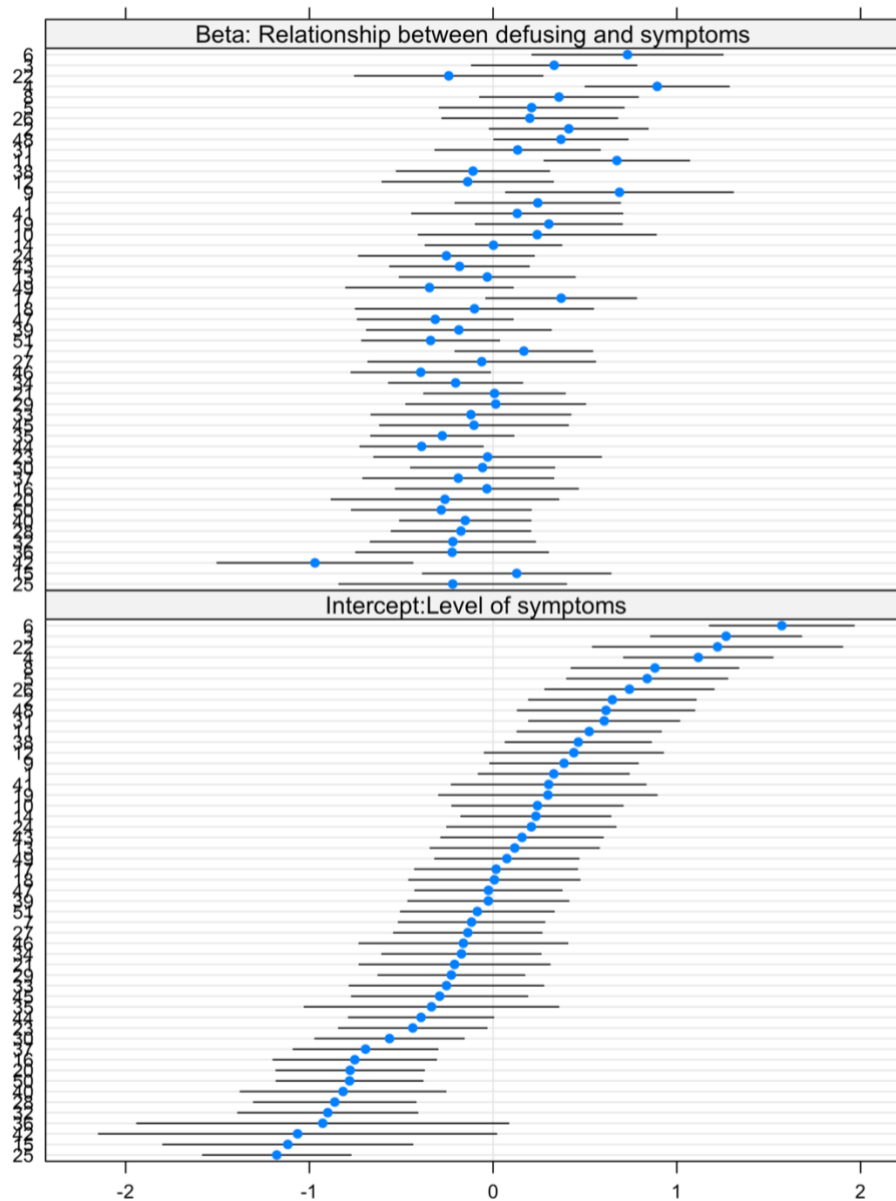


Fig. 4 Standardized scores for the intercept and within person relationship between weekly defusing and weekly symptoms

Idiomatic Approach

In order to determine whether there were underlying subgroups of change trajectories, data were submitted to a clustering procedure. The results of the clustering are shown for the

first item “being present” but we ran the analyses for all of the other items as well (see supplementary section B for details). For the item “being present”, the hierarchical clustering using AGNES is shown in Figure 5 (see the clusters for the remaining items in supplementary section D). The Dendrogram demonstrates the four-cluster solution, which was chosen for theoretical reasons (see methods section). The height of the vertical lines is a measure of similarity. The shorter the line, the more similar the observations within the clusters. AGNES used the α - and β - weights as defined above and joined the most similar two clusters in each iteration. For the above-mentioned theoretical reasons, the AGNES clustering sequence was stopped at a solution with four clusters.

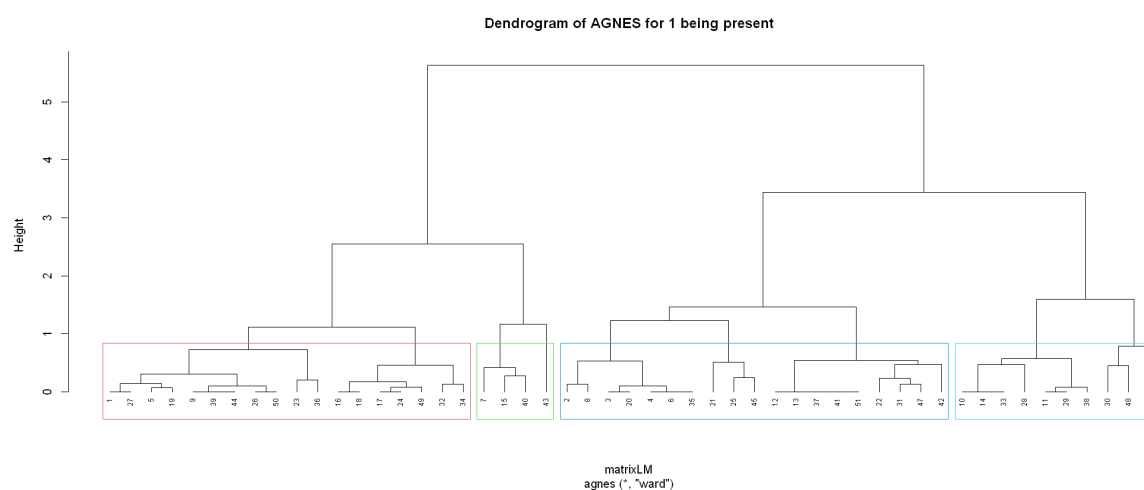


Fig. 5 Dendrogram for Item 1 (being present) using the AGNES clustering method

To better understand the outcome of the clustering sequence, we visualized the participants’ individual trend lines for “being present” grouped by the resultant four clusters (see Figure 6). The algorithm clearly divided the data into similar groups, based first on the slope (with a weight of 3) and second on the intercept (with a weight of 1). Each cluster shows a distinctive trend: patients of cluster 1 show a declining trend over the course of eight weeks, patients of cluster 2 reported little positive or no change, patients in cluster 3 reported

a large decline over the course of therapy, and patients in cluster 4 experienced a large gain in “being present”. All the other variables also produced four cluster solutions, with similar characteristics regarding their trajectories (see supplemental material section E).

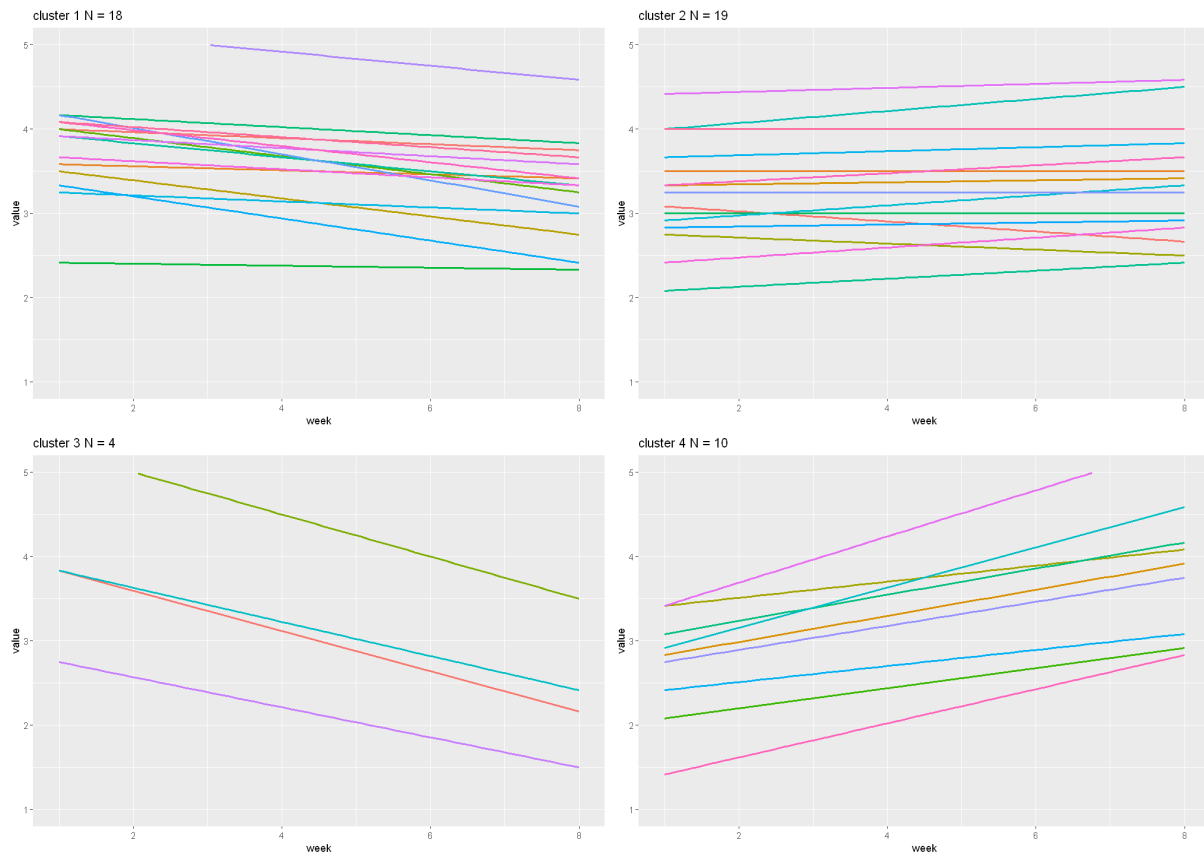


Fig. 6 Clusters of Item 1 (being present) with AGNES method

After identifying these subgroups of change, we then (as a last step instead of a first step) averaged the time series at each time point for the patients within each cluster for each variable. Figure 7 shows the data for “being present” as an example. As is clearly visible, the aggregated cluster changes differ from the average mean line. It becomes apparent that one of the largest groups (N=18) shows a slow decline with a slight gain between week 7 and 8. In contrast, there is a group (N=10) that reports continuously rising values, one small group (N=4) that reports overall declining values, and another large group (N=19) that reports no significant changes. Theoretical interpretation of this finding aside, the figure shows very

clearly that different courses exist for this variable and that clustering the sample in a meaningful way leads to interpretable results, where a single average line, where all data is collapsed together, does not. For examples of the clustering results for items 2 to 7, please refer to supplemental material section F.

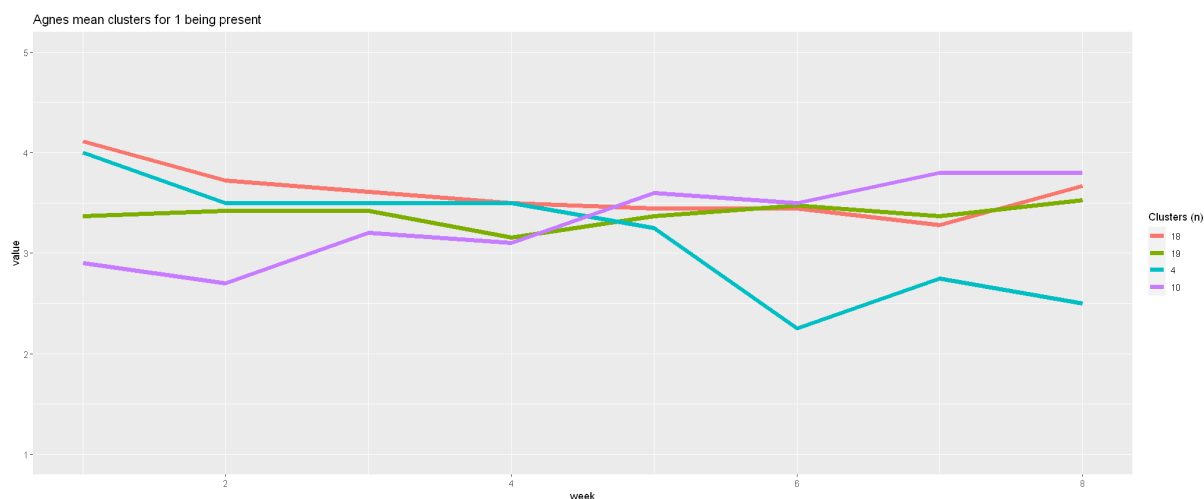


Fig. 7 Average Cluster Time Series for Item 1 (being present)

Nomothetic Prediction of Distal Outcome

In order to examine the clinical utility of these clusters, we tested whether the derived clusters per each hypothesized process items differed with respect to the distal outcome variable of well-being assessed separately at post-treatment. As expected, the omnibus ANOVA was not significant, likely because of the small sample size and the observation above that clusters cancel each other out (some patients increase, and others decrease). Based on this we examined the planned contrasts in order to understand the clinical implications of the idiomatically derived clusters. Within the idiomatically derived groups for *Own Values*, the increasing well-being scores of individuals from group 2 (relatively no change in *Own Values*, $p = .031$), group 3 (continuous gains in *Own Values*, $p = .012$), and group 4 (early gains in *Own Values*, $p = .047$) differed from group 1 (decrease in *Own Values*), with an

overall large effect size (Cohen's $f = 0.41$). For *Defusion*, the changes in well-being scores of group 4 (continuous gains in *Defusion*, $p = 0.48$) differed from group 1 (decrease in *Defusion*), resulting in a medium effect size (Cohen's $f = 0.34$). For *Acceptance*, group 3 (continuous gains in *Acceptance*, $p = .042$) and group 4 (no change in high baseline of *Acceptance*, $p = .042$) differed from group 1 (decrease in *Acceptance*), reaching a medium effect size (Cohen's $f = 0.37$). Cohen's f therefore differentially predicted well-being with respect to the process variables of *Own Values*, *Defusion*, and *Acceptance*. Although visual inspection of the graphs showed differences across the clusters, the remaining process variables showed only low effects with respect to the difference in outcomes of well-being: [*Being Present* ($f = 0.19$), *Being Engaged* ($f = 0.13$), *Steady Self* ($f = 0.13$)]. Idiomically derived clusters based on the intermediate outcome variable of *Symptoms* ($f = 0.26$) also did not result in significant differences with respect to the distal outcome of well-being, with a medium effect size. See supplementary material in section G and H for more details.

Discussion

The efficacy of psychotherapy for mental disorders is well documented. The vast majority of studies show that on average patients improve to a statistically significant degree and that this change is clinically meaningful (Butler et al., 2006; Gloster et al., 2020a; Gloster et al., 2020b; S. Hofmann et al., 2010). Although it is an oversimplification to assume that results from randomized trials apply to a given individual, this interpretation is common and individual patient trajectories are extended from such studies into clinical care. The present study presented a method for identifying different patterns of change and illustrated that meaningful subgroups can be found even when the intervention has been shown to be effective at the group level. The present study also points towards one possible way to methodologically and empirically explore such sub-groups.

Ergodicity requires both stationarity and that the same dynamic model applies to all constituent elements. In the present study, the data did not meet the ergodic assumption. We found that, for example, defusion was related with either more symptoms, less symptoms, or not related at all – depending on the individual. Furthermore, very different trajectories were present over a treatment of eight weeks. Cluster analysis based on idiographic data identified distinguishable patterns of deterioration, no change, or improvement. When looking at week-by-week patterns, a group of sudden gains was visible. When results were examined idiographically relevant to intra-individual variability, the relationship of a given process to outcome could vary from significantly positive to significantly negative. Thus, both of the key requirements for ergodicity (stationarity and common dynamic models) were violated. Because the ergodic assumption was demonstrably violated in these clinical data, relying exclusively on nomothetic results would misrepresent some individuals. It is not hard to see why from the present data: at the purely nomothetic level these idiographic differences largely cancelled each other out, and process to outcome relationships were no longer evident.

Recent studies on ergodicity have reached the same conclusion, though the percentage of a sample that belongs to a sub-group that is not characterized by the nomothetic finding differs (Gloster et al., 2023; Fisher et al., 2018). The present results also suggest one possible way forward. In the present approach, analyses began by examining entirely idiographic data linked to intra-individual variability. Nomothetic clustering procedures that avoided inter-participant averaging as input then identified multiple sub-groups of patients with varied patterns of change. These bottom-up analyses allowed for the identification of differentiated patterns person by person that predicted a clinically useful distal outcome, well-being, for at least some of the key processes (acceptance, defusion, and own values) that were targeted in the underlying psychotherapy trial. Conversely, the three other targeted processes (being

present, steady self, and being engaged) and idiomically identified patterns of symptom change, were not related to distally assessed well-being. Said simply, bottom-up idiomatic analyses lead to different conclusions than top-down nomothetic analyses. It is not widely appreciated that purely nomothetic conclusions and those based on idiographic data within the same population can be virtual opposites. This is known as *Simpson's paradox* and it is quite common in behavioral and medical areas (Kievit et al., 2013). As a simple example, in all large groups of people better typists will both type fast and make fewer errors per typed word, resulting in a negative correlation between speed and errors; nevertheless, as individuals, every single typist regardless of expertise will make more errors the faster they type, reversing that relation. Keivit et al. (2013) suggest several steps to avoid falling prey to Simpson's paradox, including several of the steps taken here: intervene; study change over time idiographically; conduct cluster analysis; and visualize the data.

In order to make progress towards personalized interventions, it is important to understand multiple patterns of change in their own right. By extension, this will help advance our understanding of the underlying processes of change initiated by an intervention. The exact number of subgroups and trajectories is likely a function of the subject matter, intervention, and clinical presentation of patients. Although these patterns shown here are largely consistent across the variables examined in this study and with patterns hypothesized in the literature (Lutz et al., 2013), we do not claim that these are the only patterns that exist. Indeed, much more research is needed to help identify the nature of change itself. We do argue, however, that the idiomatic results from the present study can be extended to individuals if idiographic patterns are the basis for such an extension.

In the present study with patients suffering from mental disorders, the idiomically derived clusters of change based on being accepting and defusing from one's thoughts as well as engaging in values were most predictive of the distal outcome of well-being. Clusters of

change patterns based on other variables were less differentially predictive of the distal outcome. Importantly, across all process variables and symptoms, the trajectory of the cluster of patients that declined the most by the end of treatment was already visually discernible from the other cluster of patients by the fourth or fifth week in the graphs depicting week-by-week change per cluster. In a fully idionomic approach, the idiographic implications of such nomothetic extensions would be formally tested. For example, when treating patients suffering from mental health issues a clinician might pay particular attention to the weekly development of acceptance, defusion and values after a month or so of treatment. If the patterns observed here are replicated in other studies, it would enable clinicians to begin to pinpoint the timepoint within an ongoing psychotherapy at which their current patients' outcome is most predictive of an outcome similar to the bottom-up derived empirical clusters. If one's current patient's course is most identified with a cluster associated with high levels on the desired distal outcome, then the therapist can hold the course. If, however, the patient's course is most similar to a cluster associated with poor prognosis, changes can be introduced. Such individualization of the treatment would save resources and hopefully improve the lives of patients via more efficient care.

It is noteworthy that these results were obtained in a real-world clinical sample that was relatively small in terms of participants and time points. The bottom-up idiographic approach thus appears to be reasonably sensitive, perhaps in part because the building of homogeneous subgroups partitions the variance in ways that increase statistical power. This could be an advantage for clinical research aimed at understanding processes of change. Other lines of research on the science of change suggest that there is fertile ground for idiographic approaches. Personalized medicine, for instance, aims to tailor interventions by grouping patients into categories based on indicators that are predictive of outcome. Within the field of psychotherapy, efforts are underway to systematically organize the understanding

of processes of therapeutic change as something separate from the techniques or therapies that give rise to change. Examining the processes of change using idionomic bottom-up approach may help realize this goal.

This study has multiple limitations. First, our sample size was relatively small. Larger sample sizes will allow us to detect more complex patterns of change and will be more representative of the clusters we are likely to identify in the population. Second, the iterative process used in this study of calculating the clusters separately for each variable does not allow for the real possibility that the variables dynamically influence each other's development. We did not have sufficient power to conduct such dynamic analysis. To our knowledge, such multivariate approaches are currently missing and statistical work in this area is needed. Third, although this approach allowed for clustering of shorter time series than the currently most used models, a good approach to model individual cluster time-series analysis for shorter time-series is also needed. Likewise, the idionomic approach used here relied on distance between observations to derive clusters. Approaches are needed that can also use time as a weighted parameter. Fourth, for clarity and simplicity we included participants with complete data sets. Future research may examine what effect missing values have on the results. Finally, these real-world data were derived using one intervention (i.e., Acceptance and Commitment Therapy) and the variables targeted the hypothesized processes of change. It is necessary to test whether these or other clusters of change emerge with different therapies and targeted processes of change.

To rectify these limitations, future studies should focus on systematically investigating how to translate results from this approach into clinical practice, comparing them to other emerging idionomic approaches (Gloster & Karekla, 2020; Snijders & Bosker, 2012). This could include further tests of what distal outcomes the idionomically identified

clusters predict and which they do not. Studies could also explicitly test whether tailoring treatment based on clusters improves outcomes for individual patients.

In conclusion, as clinical change applies to particular people, average is not enough. Whereas nomothetic approaches inform about the efficacy of an intervention in general, ergodicity is seldom present and the nomothetic statistics are incomplete and may at times lead to faulty conclusions. Additional, bottom-up idiomonic approaches are needed to augment traditional knowledge from clinical trials to better understand the process of change and to use this knowledge to provide better treatments for all subgroups of patients.

Declarations and Statements

Funding

This work was supported by grants from the Swiss National Science Foundation awarded to ATG (PP00P1_163716/1 & PP00P1_190082). The funder provided support in the form of salaries for authors [ATG, MN, VB, EH], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethics approval

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethikkommission Nordwest- und Zentralschweiz (EKNZ-165/13).

Consent

Written informed consent was obtained from all individual participants included in the study.

Data availability

The datasets generated during and/or analyzed during the current study are available on OSF, https://osf.io/a9hqe/?view_only=c757ab908ea94f4e99223d259096012f.

Authors' contribution

Andrew Gloster conceived and designed the study and obtained funding. Material preparation and data curation were performed by Jeanette Villanueva, Victoria Block & Charles Benoy. Data analysis was conducted by Matthias Nadler, Joseph Ciarrochi, and Victoria Block. The first draft of the manuscript was written by Andrew Gloster, Victoria Block, Steven Hayes, Joseph Ciarrochi, Elisa Haller, and Matthias Nadler. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Acknowledgments

We wish to thank all patients, who participated in this study, all dedicated treatment teams, and students who helped with the data collection and manuscript preparation.

References

- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
<https://doi.org/10.18637/jss.v067.i01>
- Birkhoff, G. D. (1931). Proof of the Ergodic Theorem. *Proceedings of the National Academy of Sciences*, *17*(12), 656–660. <https://doi.org/10.1073/PNAS.17.2.656>
- Boltzmann, L. (1885). Über die Eigenschaften monocyclischer und anderer damit verwandter Systeme). *Journal Fur Die Reine Und Angewandte Mathematik*, *1885*(98), 68–94.
<https://doi.org/10.1515/CRL.1885.98.68/MACHINEREADABLECITATION/RIS>
- Butler, A., Chapman, J., Forman, E., & Beck, A. (2006). The empirical status of cognitive-behavioral therapy: a review of meta-analyses. *Clinical Psychology Review*, *26*(1), 17–31. <https://doi.org/10.1016/J.CPR.2005.07.003>
- Cattell, R. B. (1952). The three basic factor-analytic research designs—their interrelations and derivatives. *Psychological Bulletin*, *49*(5), 499–520.
<https://doi.org/10.1037/H0054245>
- Cherstvy, A. G., Thapa, S., Wagner, C. E., & Metzler, R. (2019). Non-Gaussian, non-ergodic, and non-Fickian diffusion of tracers in mucin hydrogels. *Soft Matter*, *15*(12), 2526–2551. <https://doi.org/10.1039/C8SM02096E>
- Ciarrochi, J., Sahdra, B. K., Hayes, S. C., Hofmann, S. G., Sanford, B., Stanton, C., Yap, K., Fraser, M., Gates, K., Gloster, A. (2023). A personalised-approach to identifying important determinants of well-being [Preprint]. *OSF Preprints*.
- Edelbrock, C., & McLaughlin, B. (1980). Hierarchical Cluster Analysis Using Intraclass Correlations: A Mixture Model Study. *Multivariate Behavioral Research*, *15*(3), 299–318. https://doi.org/10.1207/s15327906mbr1503_5

- Fisher, A., Medaglia, J., & Jeronimus, B. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6106–E6115.
<https://doi.org/10.1073/PNAS.1711978115>
- Galenko, P. K., & Jou, D. (2019). Rapid solidification as non-ergodic phenomenon. *Physics Reports*, *818*, 1–70. <https://doi.org/10.1016/j.physrep.2019.06.002>
- Gates, K. M., Chow, S.-M., & Molenaar, P. C. M. (n.d.). *Analysis of intra-individual variation: Systems approach to human processing*. Chapman & Hall/CRC.
- Gates, K. M., Lane, S. T., Varangis, E., Giovanello, K., & Guiskewicz, K. (2017). Unsupervised Classification During Time-Series Model Building. *Multivariate Behavioral Research*, *52*(2), 129–148. <https://doi.org/10.1080/00273171.2016.1256187>
- Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319. <https://doi.org/10.1016/J.NEUROIMAGE.2012.06.026>
- Gloster, A. T., Block, V. J., Klotsche, J., Villanueva, J., Rinner, M. T. B., Benoy, C., Walter, M., Karekla, M., & Bader, K. (2021). Psy-Flex: A contextually sensitive measure of psychological flexibility. *Journal of Contextual Behavioral Science*, *22*, 13–23.
<https://doi.org/10.1016/J.JCBS.2021.09.001>
- Gloster, A. T., Haller, E., Villanueva, J., Block, V., Benoy, C., Meyer, A. H., Brögli, S., Kuhweide, V., Karekla, M., Bader, K., Walter, M., & Undine, L. (2023). *Transdiagnostic Therapy for Non-Responding In- and Outpatients with Common Mental Disorders: The Nonrandomized “Choose Change” Effectiveness Trial*.
<https://doi.org/10.1159/000529411>
- Gloster, A. T., & Karekla, M. (2020). A multilevel, multimethod approach to testing and refining intervention targets. In S. C. Hayes & S. G. Hofmann (Eds.), *Beyond the DSM:*

Toward a process-based alternative for diagnosis and mental health treatment (pp. 225–249). Context Press/New Harbinger Publications.

Gloster, A. T., Klotsche, J., Ciarrochi, J., Eifert, G., Sonntag, R., Wittchen, H. U., & Hoyer, J. (2017). Increasing valued behaviors precedes reduction in suffering: Findings from a randomized controlled trial using ACT. *Behaviour Research and Therapy, 91*, 64–71.
<https://doi.org/10.1016/j.brat.2017.01.013>

Gloster, A. T., Klotsche, J., Gerlach, A. L., Hamm, A., Ströhle, A., Gauggel, S., Kircher, T., Alpers, G. W., Deckert, J., & Wittchen, H. U. (2014). Timing matters: Change depends on the stage of treatment in cognitive behavioral therapy for panic disorder with agoraphobia. *Journal of Consulting and Clinical Psychology, 82*(1), 141–153.
<https://doi.org/10.1037/a0034555>

Gloster, A. T., Rinner, M., Ioannou, M., Villanueva, J., Block, V., Ferrari, G., Benoy, C., Bader, K., & Karekla, M. (2020a). Treating treatment non-responders: A meta-analysis of randomized controlled psychotherapy trials. *Clinical Psychology Review, 75*.
<https://doi.org/10.1016/J.CPR.2019.101810>

Gloster, A. T., Walder, N., Levin, M. E., Twohig, M. P., & Karekla, M. (2020b). The empirical status of acceptance and commitment therapy: A review of meta-analyses. *Journal of Contextual Behavioral Science, 18*, 181–192.
<https://doi.org/10.1016/J.JCBS.2020.09.009>

Gonzalez, R., & Palais, S. (1962). *A path-building procedure for iterative circuit computers*.
<https://doi.org/hdl.handle.net/2027.42/5167>

Gottlieb, G. (1992). *Individual Development and Evolution: The Genesis of Novel Behavior*. Oxford University Press.

- Hayes, S. C., & Hofmann, S. G. (2021). “Third-wave” cognitive and behavioral therapies and the emergence of a process-based approach to intervention in psychiatry. *World Psychiatry, 20*(3), 363–375. <https://doi.org/10.1002/WPS.20884>
- Hayes, S. C., Hofmann, S. G., & Ciarrochi, J. (2020). A process-based approach to psychological diagnosis and treatment: The conceptual and treatment utility of an extended evolutionary meta model. *Clinical Psychology Review, 82*, 101908. <https://doi.org/10.1016/J.CPR.2020.101908>
- 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*(April), 40–53. <https://doi.org/10.1016/j.brat.2018.10.005>
- Hofmann, S. G., & Hayes, S. C. (2019). The Future of Intervention Science: Process-Based Therapy. *Clinical Psychological Science, 7*(1), 37–50. <https://doi.org/10.1177/2167702618772296>
- Hofmann, S., Sawyer, A., & Fang, A. (2010). The empirical status of the “new wave” of cognitive behavioral therapy. *The Psychiatric Clinics of North America, 33*(3), 701–710. <https://doi.org/10.1016/J.PSC.2010.04.006>
- Horst, U. (2008). Ergodicity and Nonergodicity in Economics. In *The New Palgrave Dictionary of Economics* (pp. 1–6). Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95121-5_2498-1
- Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University*. <https://doi.org/10.13140/RG.2.2.11388.90240>

- Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review, 30*(4), 865–878.
<https://doi.org/10.1016/j.cpr.2010.03.001>
- Kelderman, H., & Molenaar, P. C. M. (2007). The Effect of Individual Differences in Factor Loadings on the Standard Factor Model. *Multivariate Behavioral Research, 42*(3), 435–456. <https://doi.org/10.1080/00273170701382997>
- Keyes, C. L. M., Wissing, M., Potgieter, J. P., Temane, M., Kruger, A., & van Rooy, S. (2008). Evaluation of the mental health continuum-short form (MHC-SF) in Setswana-speaking South Africans. *Clinical Psychology and Psychotherapy, 15*(3), 181–192.
<https://doi.org/10.1002/cpp.572>
- Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology, 4*(AUG), 513. <https://doi.org/10.3389/FPSYG.2013.00513/BIBTEX>
- Kuehn, N. M., & Abrahamson, N. A. (2020). Spatial correlations of ground motion for non-ergodic seismic hazard analysis. *Earthquake Engineering and Structural Dynamics, 49*(1), 4–23. <https://doi.org/10.1002/eqe.3221>
- Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M., 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 : Journal of the Society for Psychotherapy Research, 23*(1), 14–24.
<https://doi.org/10.1080/10503307.2012.693837>
- Miller, M., & Van Horn, J. (2007). Individual variability in brain activations associated with episodic retrieval: a role for large-scale databases. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology, 63*(2), 205–213. <https://doi.org/10.1016/J.IJPSYCHO.2006.03.019>

- Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research and Perspectives*, 2(4), 201–218.
https://doi.org/10.1207/s15366359mea0204_1
- Molenaar, P. C. M. (2013). On the necessity to use person-specific data analysis approaches in psychology. *European Journal of Developmental Psychology*, 10(1), 29–39.
<https://doi.org/10.1080/17405629.2012.747435>
- Molenaar, P. C. M., & Campbell, C. G. (2009). The New Person-Specific Paradigm in Psychology. *Current Directions in Psychological Science*, 18(2), 112–117.
<https://doi.org/10.1111/J.1467-8721.2009.01619.X>
- Peters, H. N., & Sidman, M. (1961). Tactics of Scientific Research: Evaluating Experimental Data in Psychology. *The American Journal of Psychology*, 74(3), 484.
<https://doi.org/10.2307/1419766>
- Sahdra, B. K., Ciarrochi, J., Ferrari, M., Yap, K., Haller, E., Hayes, S. C., Hofmann, S. G., Gloster, A. (2022). There is no self without the other: An experience sampling study of the interconnectedness of self-and other-compassion and its link to well-being [Preprint]. *PsyArXiv*.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications.
- von Neumann, J. (1932). Proof of the Quasi-Ergodic Hypothesis. *Proceedings of the National Academy of Sciences*, 18(1), 70–82. <https://doi.org/10.1073/PNAS.18.1.70>
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), 236–244.
<https://doi.org/10.1080/01621459.1963.10500845>
- Wohlwill, J. F. (1973). *The study of behavioral development*. Academic Press.

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2(2), 165–193. <https://doi.org/10.1007/S40745-015-0040-1>