Title

Bringing preference into precision psychiatry: A forced-choice measure of preferences for symptom change in depression.

Description

The question how to efficiently assign patients to the optimal treatments, lies at the heart of much medical research. In psychiatry, approaches fall on opposite ends of a decision-making continuum from clinician-led to patient-led (Kon, 2010): On one end, 'precision psychiatry' has been seeking aetiological factors to create better algorithms for assigning patients to treatments - While precision psychiatry sets its scope from physiology to demography (Fernandes et al., 2017), it seems intended largely without consideration of what a particular patient 'wants' or 'prefers'. On the other side, a canopy of tools has been developed to measure and accommodate patients' preferences (Swift et al., 2018), but such approaches have remained largely agnostic to symptomatology and aetiology. However, in mood disorders, preference and symptom are intricately interwoven (Sheppes et al., 2015; Vanderlind et al., 2021) - thus, for many patients, true precision psychiatry will happen right in between. But, to find out, we need the right tools: The goal of this study is to contribute towards bridging the gap between aetiology and preference by validating a new tool which assesses patients' preference for symptom change - that is, to identify which symptoms they would like to see changed most urgently.

In meta-analysis, accommodation of patient preference improves psychotherapeutic outcomes (d = 0.29; Swift et al., 2018). However, so far preference research has remained largely separate from the precision-psychiatric paradigm. In line with Swift et al. (2018), preferences have often been defined as 'the specific conditions and activities that clients want in their treatment experience", which may include preferences for specific treatment modalities and attributes (e.g., Lokkerbol et al., 2018; Sandell et al., 2011), or activities during treatment (Cooper et al., 2020; Cooper & Norcross, 2016). In this vein, preference is largely considered an individual difference without clear aetiological origins or implications - thus, preference may seem relevant to day-to-day decision-making, but not to a precision-psychiatry framework.

However, recent evidence indicate that precision psychiatry still requires an understanding of patients' preferences - albeit preference of a slightly different nature. What a patient prefers or not can often be linked to the aetiology and symptomatology of their disorder. Conceptually, emotion regulation is a motivated process (Tamir, 2016, 2020) - that is, I may change my emotion regulation strategies in line with the perceived value of emotional states (and their utilitarian consequences). Athletes who believe anxiety helps their performance report up-regulating their anxiety before a competition (Lane et al., 2011), and negotiators who think anger helps them to assert themselves seek out more anger-inducing stimulus material before a confrontation (Tamir & Ford, 2012).

Importantly, emotion preferences are implicated in depressive (Vanderlind et al., 2019) and anxiety-related (Vanderlind et al., 2022) symptomatology. Depressed participants are more likely to seek out low-mood inducing stimuli (Millgram et al., 2015) and are more likely to consider low-mood inducing activities such as rumination as useful (Watkins & Baracaia, 2001). Similarly, participants with low self-esteem are less likely to repair induced negative mood states (Wood et al., 2008). A longitudinal study (Millgram et al., 2018) found that reduced preference for happiness predicted an increase in depressive symptoms in college students 6 months later. In conclusion, preference for emotion and depressive symptoms overlap - indeed, emotion preference may be implicated in the aetiology of mood disorders (Vanderlind et al., 2021). Thus, a precision psychiatric account of mood disorders will have to account for variability in how participants value their own emotions, and hence, their emotion-related symptoms.

In this study, we aim to provide a tool that can link symptom and preference, by asking participants which of their symptoms they would like to alleviate the most. Based on a broad battery of depression symptoms (Fried & Nesse,

2015), we propose a forced-choice measure of symptom preference. Our proposed tool generates an individualised ranking of symptoms for patients, which may in the future allow practitioners to prioritise symptoms. In contrast to Likert scales, pairwise forced decision tasks ask participants to select between two options on the basis of some prompt. In this study, we use the prompt "If you could make one of these two problems go away, which would you pick?" to determine the symptoms that individuals with depression would most like to be rid of. Starting from a list of 52 symptoms taken largely from Fried & Nesse, participants selected those that had impacted them in the previous 6 months, then we asked them to rate the frequency, severity, and impact of each of those symptoms on 6 level likert scales. We then collected pairwise forced-choice comparisons of all of the endorsed symptoms.

Forced-choice methods are common in marketing, and have been successfully used in patient preference research in the past. Notably, Lokkerbol et al. (2018), have investigated depressed patients' preferences for treatment modalities represented by vignettes - finding, e.g., that participants generally preferred face-to-face therapies over online formats, as well as shorter over longer waittimes. The advantage of forced-choice is twofold. First, forced-choice methods limit ceiling and floor-effects (e.g., Meade et al., 2004). This may be particularly useful for indecisive patients who consider most symptoms as either very pressing, or not pressing at all - Hence, exactly the group where a formalised preference-assessment tool would be most needed. Second, forced-choice measures provide an easily interpretable 'priority list' which can effectively guide clinician's decisions, or form the basis of further shared discussion.

Hypotheses and Goals:

For now, our main goal is to provide some preliminary evidence of the reliability and validity of the tool. First, we would like to establish the internal consistency (H1.1) and test-retest reliability of our tool (H1.2). Second, we aim to provide evidence for the validity of our approach. We aim to show that participants' response process during the forced-choice task is consistent with accessing, and then comparing, the preference for change assigned to each item. We assume that it is harder for participants to decide between more similarly ranked items. In line with the well-established finding that harder decisions require longer processing times (Hick, 1952, Hyman, 1953, Hu et al., 2022), we seek to test whether smaller rank differences produce longer reaction times on the forced-choice task (H2.1)
Another aspect of validity is convergent validity: The concept of 'preference for symptom change' rests on the assumption that preference to change a symptom is not the same as (but related to) the severity, frequency and impact that symptom has on daily life (H2.2). A third key question is whether there is sufficient variability in preference (H3): If all patients have the same preferences for symptom change, measuring preference in practice may not be worthwhile. Finally, we seek to explore some implications of our data for precision psychiatry: notably, what gives rise to preference for change - are symptom frequency, severity or impact more important (H4)? Are there clusters of patients with similar preferences (H5)? If so, which characteristics are associated with those clusters? Are there groups of patients that are particularly indecisive, such that formalised assessment of preference with a forced-choice task is most useful (H6)?

H1: The forced-choice method is reliable. Operationalised as:

- H1.1: Participants' responses on the forced-choice task will differ significantly from responses that would be obtained under random responding.
    - 1.1a) The number of wins in core items reflect non-random responses across participants
    - 1.1b) We plan to measure the proportion of transitively consistent preferences, a proportion over 70% would be acceptable
    - 1.1c) We plan to measure the proportion of participants whose entropic profile is significantly different from that obtained under random responding, a proportion of over 70% would be acceptable
- H1.2: Participants' responses on the forced choice task will correlate across time points
    - 1.2.a) The number of wins in core items will correlate across time points

    - 1.2.b) The Aitchison distance between participants' responses in the first and second wave will be smaller than expected under random responding.

H2: The forced-choice method has convergent validity. Operationalised as:

- H2.1: Within waves, larger rank differences incur smaller decision times during forced-choice
- H2.2: Within waves, agreement with Likert-items is positively associated with scores on the forced choice questionnaire.
    - H2.2.a) Likert items will be positively associated with preferences when correlating Likert-items with the number of wins/item ranks of core items.
    - H2.2.b) Within core-items, Likert items are positively associated with preferences in a Bradley-Terry model, individually predicting the relationship between agreement on a given item and the ability/preference for that item.
    - H2.2.c) The difference in Likert-item agreement of any two items is associated with the rank difference of those items.

H3: There will be meaningful variation in participants' preference profiles.

H4: (Exploratory) Frequency, severity, and impact will differ in the extent to which they predict preference.

H5: (Exploratory) Participants' preference profiles will form clusters, and cluster-membership will be predicted by demographic characteristics and depression.

H6: (Exploratory) Participants' decisiveness (operationalised via the amount of entropy in participants' preference profile) will correlate with demographic characteristics and depression.

Design plan

Study design

Our proposed analysis uses pre-collected data from longitudinal two-wave design, with data-collection taking place 3 days apart in 2021. On each measurement occasion exactly the same materials and procedure was used. The data was purely observational, there was no active manipulation.

Sampling Plan

Explanation of existing data

To avoid bias, we have limited researchers' access to the full data prior to pre-registration. While all authors had access to the full data, we took efforts to not engage with the data. So far, we only: (a) plotted histograms of raw wins in the forced choice task, as well as missingness statistics to determine data quality, and (b) conducted several pilot analyses on a subset of randomly chosen 20 participants.
After formulating our main hypotheses, we chose to try them out on the data prior to pre-registration - primarily to gauge the feasibility of our analysis plan. In particular, the forced choice task produces ipsative data, for which many traditional analysis methods are unsuitable. This meant that we had to pivot from some standard analyses (e.g., Cronbach's alpha) to more

suitable, but less familiar alternatives (e.g., Kendall/David's c' statistic; Mazzuchi et al., 2008). The pre-exploration ensured the quality of our analyses and suitability of the proposed methods for the data at hand.

Data collection procedures

Data from 130 US-based English-speaking adults was collected via CloudResearch, with a compensation of $17.50 per hour. The participants previously had indicated on a previous survey that they "Struggle with depression" (Yes/No item).

First, participants completed a short demographic questionnaire recording sex, age, and their score on the centre for epidemiological studies depression questionnaire (CESD). Then, participants were given a list of 52 items taken from existing questionnaires. The items were chosen so as to maximise coverage across the range of depression-relevant symptoms identified by Fried & Nesse (2015). Symptoms were taken from the mood and feelings questionnaire (MFQ, 4 items) and the centre for epidemiological studies depression questionnaire (CESD, 17 items), the quick inventory of depressive symptomatology (QIDS, 3 items), the full inventory of depressive symptomatology (IDS, 10 items), the Zung self-rating depression scale (SDS, 8 items) and the Beck-Depression Inventory (4 items). 6 additional items were included to improve coverage of depression-related symptoms listed by Fried & Nesse (2015), but that were not covered in the questionnaires (e.g., weight gain, initial/middle insomnia, anxiety, phobia, gastro-intestinal symptoms). Five symptom domains from Fried & Nesse were excluded due being unsuitable for self-report (hypochondriasis, loss of insight), management/ethical risks in online collection (suicidal ideation, decreased libido) and overlap with other content domains (inability to feel, showing strong content-overlap with loss of interest).

They were asked to choose "Which of these problems have bothered you in the previous 6 months?". Participants could select as many symptoms as they liked. In the following task, participants were shown their selected symptoms as well as 5 core symptoms, which thematically covered anhedonia or low mood (example: "Feeling depressed"). For each symptom participants endorsed, plus the 5 core symptoms regardless of endorsement, we collected Likert responses to the frequency of the symptom ("How often has this problem bothered you in the last 6 months?", scored '0 - Never','1 - Only once','2 - Sometimes','3 - Weekly','4 - Every few days','5 - Every day'), its severity ("When this problem bothers you, how severe is it?', scored: '0 - No bother','1 - Minimal','2 - Mild','3 - Moderate','4 - High','5 - Most severe') and its impact on life ("How much does this problem impact your life overall?, scored: '0 - No impact','1 - Minor annoyance','2 - Mild','3 - Troublesome','4 - High impact','5 - Prevents me from going about my life').

Next, participants completed the main forced-choice task. Participants were shown one randomly chosen item from the pool of endorsed/core items on each side of the screen. Their prompt was: "If you could make one of these two problems go away, which would you pick?". They indicated their choice via button press.

The number of comparisons each participant was shown was capped at 190. From previous unpublished forced-choice questionnaires, we anticipated a response time of 3s per item. This threshold was chosen to keep the length of the study within roughly 10 minutes. If a participant endorsed too many symptoms, a random selection of the possible comparisons was shown.

Sample size

Our data structure is a 3-level nested design, involving 130 participants, completing 2-waves of data collection, including a maximum of 190 forced-choice trials per wave per participant.

Sample size rationale

For most of our proposed analyses, no agreed-upon methods exist for computing statistical power. Given time constraints, we could not conduct extensive simulation-testing prior to data collection. Instead, our sample size was based on practical constraints. In particular, the two-wave design limited the number of participants for whom data could be collected within the available resources.

Variables

Measured variables (all wave 1 & wave 2)

- Participants' age and sex
- Depression-score on the CESD
- Selection of 'most bothersome' items on our depression questionnaire
- Likert-item responses on the severity, frequency and impact of the endorsed (+ core) symptoms
- Wins on the forced choice questionnaire (where a 'win' is allocated to an item if it is chosen over another item).
- Reaction time for selecting the winning item after the pair of items was presented.

Indices

Forced choice rankings. For each item from the forced-choice questionnaire, a ranking is constructed - at both the participant-and the sample-level. Participant-level rankings are assigned based on the number of times this item was chosen by a given participant. Sample-level rankings are derived from the number of wins this item has received in total, in the entire sample.

Composite symptom impairment measure. In order to limit co-linearity between the severity, frequency and impact ratings for each item, we plan to compute their sum as a measure of symptom-induced impairment. To determine whether this is needed, we will do the following: Pooling across items, we will conduct an exploratory factor analysis to determine the shared loadings of all three items on a shared underlying factor. If loadings are larger than 0.7 (standardised), we will use the sum score instead of using the items independently.

Composite entropy measure. See below and demonstration.qmd for details.

Analysis Plan

Statistical models (required)

H1.1. Internal consistency. Our first goal is to examine the extent to which participants' responses on the forced-choice questionnaire reflect true variance in their responses, rather than random variation. Since there is no universally agreed upon method for computing internal consistency in 'ipsative' forced choice data (e.g., Van Leeuwen & Mandabach, 2002; Dunlap & Greer, 1997), we will use a range of convergent measures - in particular, testing the null-hypothesis whether the pattern of wins each item has received is consistent with random responding, or reflects true underlying 'preference'.
First, in line with recommendations by Van Leeuwn & Mandabach (2002), we will conduct an omnibus within-subjects F-test on the number of wins each item has received, using item-identity as a categorical predictor. Since not all items are seen by all participants, we restrict ourselves to the 5 'core items' which are guaranteed to be present in all participants' data. This tests the null-hypothesis that the pattern of wins in the data is consistent with random responding.

Second, we will use the c' statistic proposed by Kendall (1962) and David (1963), and taken from Mazzuchi et al. (2008). This method assumes that an ideal responder with true preference would always choose consistently. Hence, the number of non-transitive item pairs (i.e., pairs where participants prefer A > B and B > C, but then choose C > A) provides a measure of inconsistency. The c'-statistic assesses whether the number of intransitive item triads observed in the data is statistically significant from the number of intransitive item triads that would be expected if the participant was responding at random. This test is run at the participant level, yielding a proportion of participants whose responding is significantly different from random. We will assume that a proportion of more than 70% of participants responding non-randomly according to this test indicates acceptable reliability. An illustration of how the c'-statistic can be used is provided in demonstration.qmd.

Third, we will use an information-theoretic approach (see e.g., Dimitrov et al., 2011). Entropy describes the average amount of 'surprisal' associated with the outcome of a random experiment. Surprisal would be low for a highly expected outcome (e.g., not winning in the lottery) whereas it is high for an unexpected outcome (e.g., winning the lottery). It is known that for discrete multi-outcome random variables such as the number of wins each item achieves, the uniform distribution maximises entropy. A uniform distribution of wins would also be observed if a participant is choosing perfectly at random. For this reason, entropy provides a quantifiable measure of how 'random' participants' responding is, with lower values of entropy representing less randomness. Consequently, we will test the null hypothesis that the entropy of participants' response profile is smaller than what would be expected under random responding. For each participant, we will compute the total entropy of the distribution of forced-choice wins. Then, we will run a simulation in which an agent repeatedly assigns wins to items at random. Based on 1000 randomly simulated samples, we will estimate the lower 5% quantile for entropy. The null hypothesis will be rejected if the entropy computed for the actual participant falls below this 5% quantile. This is equivalent to conducting a one-sided permutation test. This test is run at the participant level, yielding a proportion of participants, whose responding is significantly different from random. We will assume that a proportion of more than 70% of participants responding non-randomly according to this test indicates acceptable reliability (the expected false-positive rate would be 5%). This approach is illustrated in demonstration.qmd. We will also use the entropy-analysis to determine an entropy score for each participant, defined as one minus the proportion of theoretically attainable entropy that was actually observed. This score indicates 'decisiveness', such that higher scores show a participant who is more 'decisive'.

We propose to use multiple convergent methods to assess reliability. To understand convergence between the two entropy-analysis and the c'-test, we will provide a 2x2 count table, indicating agreement in classification between both methods. Data quality in cases where disagreement occurred will be visually examined through scatterplots, and potential adjustments (e.g., exclusion of outliers) may be made. Then, we will recompute an overall reliability-proportion, defined as the percentage of participants for whom the c'-test *or* the entropy-analysis indicate significantly non-random responding. We would require that this metric satisfies the same criterion as the individual tests: that the proportion of 'false responders' does not exceed 70% of the sample.

H1.2. Test-retest reliability. Next, we will examine test-retest reliability. This will be achieved through two methods. For core symptoms, we will compute the correlation between the number of wins assigned to this symptom in the first wave and the second wave. Pearson's correlation will be used, unless visual examination of scatterplots suggests violation of assumptions. In this case, Spearman's correlation will be employed. Standard correlation methods are normally not suitable for forced-choice data due to a sum constraint (the number of 'wins' cannot exceed the number of trials). However, in this case, only a subset of items (the core-items) is used. Hence, the total sum of wins distributed to the core items is no longer constrains the degrees of freedom and correlations can be used (Van Leeuwen & Mandabach, 2002). We will examine the average correlation across all items. We will judge a correlation point-estimate of at least

0.7 or greater as acceptable To account for uncertainty in the point estimate, we will furthermore test the null hypothesis that the test-retest reliability is significantly larger than 0.6 using confidence intervals.
To enable inference on the entire dataset, Aitchison's distance (see van Eijnatten et al., 2015) will be computed between the preference-pattern from wave 1 and wave 2, individually for each participant. Aitchison's distance quantifies the similarity between two sets of compositional data (i.e., data satisfying a sum-constraint), where 0 corresponds to identity. Thus, we aim to test the null-hypothesis that Aitchison's distance between wave 1 and wave 2 is smaller (i.e., closer to identity) than would be expected at random. Consequently, a permutation distribution will be constructed by randomly permuting the wins assigned to wave 2 for 1000 iterations for each participant. The null-hypothesis will be rejected if, for that participant, Aitchison's distance is smaller than the lower 5% quantile of the permuted distribution. This corresponds to a one-sided hypothesis test. Again, the analysis will be run individually for each participant, and we will assume that significant results for more than 70% of participants indicate acceptable reliability. A demonstration of this method is provided in demonstration.qmd
Here, we use complementary analyses with complementary weaknesses. For this reason, we will interpret each sub-analysis separately, and carefully investigate reasons for any discrepancies - keeping the strengths and weaknesses of the methods in mind.

H2.1. Rank difference and reaction time. Next, we will test whether participants respond faster on 'easier' trials (i.e., trials with a larger rank difference between items). For this purpose, participant-level ranks will be computed for each item and for each trial the absolute rank difference between the two items will be computed. Log-reaction time will be used to account for skew in reaction time. A two-level linear mixed effects will be used, using rank difference as independent variable, reaction time as dependent variable and participant as a grouping factor. First, a model without random effects will be estimated. Then, random intercepts, and finally random slopes, will be included. Model fit will be compared using log-likelihood tests and the Aikaike information criterion. The hypothesis will be evaluated by checking the significance of the rank_difference term. Due to lower missingness, the analysis will be computed first on wave 1 and then replicated on wave 2.The maximal model would look as follows in R-style pseudocode:

Log_rt ~ rank_difference + (1 + rank_difference | participant)

H2.2. Rank and agreement with Likert items. Due to the special structure of the data, multiple convergent analyses will be run to examine agreement with Likert items.
First, only core items will be analysed. Ignoring item-identity, a 2-level linear mixed effects model will be used to predict the number of wins an item has obtained (dependent variable) from the Likert-rating assigned to that item (independent variable). The model will first be fitted without random effects, then inclusion of random intercepts and random slopes will be tested using likelihood ratio tests and the AIC. The maximal model would be the following, where our hypothesis is evaluated using the significance of the Likert-agreement term.

(1) n_wins ~ likert_agreement + (1 + likert_agreement | participant)

However, the raw number of wins may be biased in some participants, since the number of times each item was presented to each participant (and hence the number of opportunities for that item to 'win') differed randomly across participants. For this reason, we plan to run an extended Bradley-Terry model (Bradley & Terry, 1952; Firth, 2006) including response to the Likert items as predictors. The Bradley-Terry Model will be coded using the BradleyTerry2 package for R (Turner & Firth, 2012). In the formalism used by Turner & Firth (2012), the logistic model predicting item ability will be represented as follows. Note that the last item will be left out as a reference item. Since the BradleyTerry2 package cannot deal with missingness in judge-level predictors, only core items are considered. This analysis will be conducted on the sum score of frequency, severity and impact. The hypothesis would be evaluated

separately for each item using the corresponding interaction terms. Due to lower missingness, the analysis will be computed first on wave 1 and then replicated on wave 2.

(2) ability = item + filter1[participant] * item1[..] + filter2[participant] * item2[..] + ….

To cover all non-core items as well, we will assess the association of participant-level rank difference between two items with the difference between the filter items assigned to this participant. The maximal model would look as follows, using the same methods of model comparison as above. Again, we would test the hypothesis by testing the significance of the likert_difference parameter:

(3) rank_difference ~ likert_difference + (1 + filter_difference | participant)

Here, we use complementary analyses with complementary weaknesses. For this reason, we will interpret each sub-analysis separately, and investigate reasons for any discrepancies, keeping the strengths and weaknesses of the methods in mind.

H3. Uniqueness of preference. We want to examine the extent to which participants' preference profiles were unique. This was achieved in two ways.
First, for the core symptoms, we randomly group participants into pairs. Then, Aitchison's distance will be computed between the participant-level rankings of all symptoms shared between the two participants. Recall that an Aitchison distance of 0 is equivalent to identity. For this reason, testing whether Aitchison's distance between two distributions is larger than 0 is equivalent to testing the null hypothesis that the two distributions are the same. We will collect the Aitchison distance for each of the couples. Next, an independent-samples t-test or non-parametric equivalent will be conducted to test whether the mean Aitchison distance between the pairs is larger than 0. Due to lower missingness in wave 1, the analysis will be computed first on wave 1 and then replicated on wave 2.
Second, for the full dataset, we will proceed by visual examination of bar-plots showing the number of wins individual participants have assigned across items. An amount of meaningful variation in preference pattern between participants, as judged by the authors, will be taken as evidence that participants' preference profiles do differ. Upon publication, the corresponding plots will be offered to the reader for their own judgement. To avoid selection bias, the set of participants destined to be plotted in case of publication will be determined pseudo-randomly. Visual examination will take place separately in both waves.

Inference criteria
A significance threshold of alpha = 0.05 will be used. For the Bradley-Terry-Model, Bonferroni-correction for multiple testing will be applied, correcting for the number of parameters in the model whose significance tests are interpreted (the actual number of parameters estimated may be larger due to computational constraints of the BradleyTerry2 package). Model fit will be evaluated using log-likelihood ratio tests, AIC and BIC. In case of conflict, we will follow the conclusion that at least ⅔ of indices indicate. For the reliability analyses relying on simulation/permutation, the significance threshold will be the corresponding (one-sided) 5%-quantile of the permuted/simulated population distribution, for each participant without multiple-comparison correction. We will judge a rate of at least 70% of significant participants as meaningfully non-random at the sample level. We arrived at this threshold by looking at the pilot data, finding that 20/20 participants had significantly non-random responding on the c' statistic, and 14/20 had significantly non-random responding on the entropy test in wave 1 (see further justification in demonstration.qmd). For analyses involving only a single timepoint, the analyses will be run first on wave 1 and then be replicated in wave 2. If there are discrepant conclusions, both models will be offered to readers, and potential reasons will be discussed.

Data exclusion & Missing Data
Where needed, we will exclude data on a trial-level. This includes missingness in the forced-choice response and unusually fast (< 300ms) and unusually slow (> 10s) responses to the forced-choice questionnaire. We do not plan to impute missing

data. Incomplete datasets will be used where possible. Other types of exclusions may be employed where necessary. We endeavour to report reasons for all such exclusions in the final manuscript. Participants who only completed one wave will be included for all analyses except the test-retest reliability analysis.

To preserve statistical power, all further analyses will be initially run on the full sample, but we will conduct a sensitivity analysis, replicating our results after exclusion of participants who failed both the transitivity and entropy-tests for internal consistency in the wave in which the analysis is run. If there is no disagreement between both analyses on significance and directionality of the effects under examination, parameter estimates from the full model will be reported. Otherwise, we will report and interpret parameters from the model without 'random responders.' For the exploratory cluster analysis (see below), we plan to initially include the whole sample but validate in the reduced sample.

Exploratory analyses

H4 Importance of frequency. We expect that severity, frequency and impact of symptoms are differentially associated with participants' preferences. However, we have no prior hypothesis about which of the three may be most important. To test this, we will run the linear mixed effects model specified under H2.2 separately using severity, frequency and impact as dependent variables. We will compare model fit using the AIC and BIC to determine if there is a meaningful difference in their effect on participants preference for change.

H5 Clustering. We predict that participants' preference profiles may form clusters. We will use k-means clustering to group participants' preferences (based on Likert-responses or raw item wins) into clusters. First, we will apply a version of multi-dimensional scaling to reduce the number of variables to a manageable level. We plan to extract 2 or 3 dimensions to maintain interpretability, but different solutions will be computed and compared using AIC and BIC. For k-means clustering, the number of clusters will be determined using the elbow method (visual examination of a scree-like plot, showing the drop in mean within-cluster variance as the number of clusters is increased), and validated using the silhouette method. If more than one cluster is found, we will then conduct a range of exploratory analyses. First, we will visually examine the preference profiles giving rise to the clusters and compute a Bradley-Terry-Model using cluster-membership as predictor. Next, we would like to test for association between demographic variables (age, sex, depression-severity) and derived variables (entropy score) and cluster-membership.

H6 Decisiveness. Based on the entropy-analysis reported above, we plan to compute an 'entropy score', computed as one minus the quotient of observed entropy in a participants' preference profile and the maximal entropy that can be attained (i.e., uniformity). Conceptually, this gives a measure of 'decisiveness' - where higher scores indicate that participants had more pronounced preferences. We would like to explore whether participants with larger depression scores report lower entropy scores (i.e., are less decisive on the questionnaire).

Reference List:

Cooper, M., van Rijn, B., Chryssafidou, E., & Stiles, W. B. (2021). Activity preferences in psychotherapy: what do patients want and how does this relate to outcomes and alliance? *Counselling Psychology Quarterly*, *35*(3), 1–24. https://doi.org/10.1080/09515070.2021.1877620

Cooper, M., & Norcross, J. C. (2016). A brief, multidimensional measure of clients' therapy preferences: The Cooper-Norcross Inventory of Preferences (C-NIP). *International Journal of Clinical and Health Psychology*, *16*(1), 87–98. https://doi.org/10.1016/j.ijchp.2015.08.003

David, H. A. (1963). The method of paired comparison. In Office of Ordonance Research (Ed.), *Proceedings of the fifth conference on the design of experiments in army research development and testing*. US Army.

Dimitrov, A. G., Lazar, A. A., & Victor, J. D. (2011). Information theory in neuroscience. *Journal of Computational Neuroscience*, *30*(1), 1–5. https://doi.org/10.1007/s10827-011-0314-3

Firth, D. (2005). Bradley-Terry Models in R. *Journal of Statistical Software*, *12*(1). https://doi.org/10.18637/jss.v012.i01

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1). https://doi.org/10.1186/s12916-015-0325-4

Hick, W. E. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26. https://doi.org/10.1080/17470215208416600

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188–196. https://doi.org/10.1037/h0056940

Hu, L., Pan, X., Ding, S., & Kang, R. (2022). Human Decision Time in Uncertain Binary Choice. *Symmetry*, *14*(2), 201–201. https://doi.org/10.3390/sym14020201

Kendall, M. G. (1962). Ranks and measures. *Biometrika*, *49*(1/2), 133–137. https://doi.org/10.2307/2333473

Kon, A. A. (2010). The Shared Decision-Making Continuum. *JAMA*, *304*(8), 903–904. https://doi.org/10.1001/jama.2010.1208

Lane, A. M., Beedie, C. J., Devonport, T. J., & Stanley, D. M. (2011). Instrumental emotion regulation in sport: relationships between beliefs about emotion and emotion regulation strategies used by athletes. *Scandinavian Journal of Medicine & Science in Sports*, *21*(6), e445–e451. https://doi.org/10.1111/j.1600-0838.2011.01364.x

Lokkerbol, J., Geomini, A., van Voorthuijsen, J., van Straten, A., Tiemens, B., Smit, F., Risseeuw, A., & Hiligsmann, M. (2018). A discrete-choice experiment to assess treatment modality preferences of

patients with depression. *Journal of Medical Economics*, *22*(2), 178–186. https://doi.org/10.1080/13696998.2018.1555404

Mazzuchi, T. A., Linzey, W. G., & Bruning, A. (2008). A paired comparison experiment for gathering expert judgment for an aircraft wiring risk assessment. *Reliability Engineering & System Safety*, *93*(5), 722–731. https://doi.org/10.1016/j.ress.2007.03.011

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*(4), 531–551. https://doi.org/10.1348/0963179042596504

Millgram, Y., Joormann, J., Huppert, J. D., & Tamir, M. (2015). Sad as a Matter of Choice? Emotion-Regulation Goals in Depression. *Psychological Science*, *26*(8), 1216–1228. https://doi.org/10.1177/0956797615583295

Millgram, Y., Sheppes, G., Kalokerinos, E. K., Kuppens, P., & Tamir, M. (2019). Do the ends dictate the means in emotion regulation? *Journal of Experimental Psychology: General*, *148*(1), 80–96. https://doi.org/10.1037/xge0000477

Sandell, R., Clinton, D., Frövenholt, J., & Bragesjö, M. (2011). Credibility clusters, preferences, and helpfulness beliefs for specific forms of psychotherapy. *Psychology and Psychotherapy: Theory, Research and Practice*, *84*(4), 425–441. https://doi.org/10.1111/j.2044-8341.2010.02010.x

Sheppes, G., Suri, G., & Gross, J. J. (2015). Emotion Regulation and Psychopathology. *Annual Review of Clinical Psychology*, *11*(1), 379–405. https://doi.org/10.1146/annurev-clinpsy-032814-112739

Swift, J. K., Callahan, J. L., Cooper, M., & Parkin, S. R. (2018). The impact of accommodating client preference in psychotherapy: A meta-analysis. *Journal of Clinical Psychology*, *74*(11), 1924–1937. https://doi.org/10.1002/jclp.22680

Tamir, M., & Ford, B. Q. (2012). When feeling bad is expected to be good: Emotion regulation and outcome expectancies in social conflicts. *Emotion*, *12*(4), 807–816. https://doi.org/10.1037/a0024443

Tamir, M. (2015). Why Do People Regulate Their Emotions? A Taxonomy of Motives in Emotion Regulation. *Personality and Social Psychology Review*, *20*(3), 199–222. https://doi.org/10.1177/1088868315586325

Tamir, M., Vishkin, A., & Gutentag, T. (2020). Emotion regulation is motivated. *Emotion*, *20*(1), 115–119. https://doi.org/10.1037/emo0000635

van Eijnatten, F. M., van der Ark, L. A., & Holloway, S. S. (2014). Ipsative measurement and the analysis of organizational values: an alternative approach for data analysis. *Quality & Quantity*, *49*(2), 559–579. https://doi.org/10.1007/s11135-014-0009-8

Vanderlind, W. M., Everaert, J., Caballero, C., Cohodes, E. M., & Gee, D. G. (2021). Emotion and Emotion Preferences in Daily Life: The Role of Anxiety. *Clinical Psychological Science*, *10*(1), 109–126. https://doi.org/10.1177/21677026211009500

Vanderlind, W. M., Millgram, Y., Baskin-Sommers, A. R., Clark, M. S., & Joormann, J. (2020). Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation. *Clinical Psychology Review*, *76*(76), 101826. https://doi.org/10.1016/j.cpr.2020.101826

van Leeuwen, D. M., & Mandabach, K. H. (2002). A Note on the Reliability of Ranked Items. *Sociological Methods & Research*, *31*(1), 87–105. https://doi.org/10.1177/0049124102031001004

Watkins, E., & Baracaia, S. (2001). Why do people ruminate in dysphoric moods? *Personality and Individual Differences*, *30*(5), 723–734. https://doi.org/10.1016/s0191-8869(00)00053-2

Wood, J. V., Heimpel, S. A., Manwell, L. A., & Whittington, E. J. (2009). This mood is familiar and I don't deserve to feel better anyway: Mechanisms underlying self-esteem differences in motivation to repair sad moods. *Journal of Personality and Social Psychology*, *96*(2), 363–380. https://doi.org/10.1037/a0012881v