# Impact of response shift effects in the assessment of selfreported depression during treatment: insights from a rTMS versus Venlafaxine randomized controlled trial

Samuel Bulteau, Myriam Blanchin, Morgane Pere, Emmanuel Poulet, Jérôme Brunelin, Anne Sauvaget, Véronique Sébille

# Impact of response shift effects in the assessment of self-reported depression during treatment: insights from a rTMS versus Venlafaxine randomized controlled trial.

**Authors :**

Samuel BULTEAU[1,2], Myriam BLANCHIN[1], Morgane PERE[3], Emmanuel POULET[4,5], Jerome BRUNELIN[4], Anne SAUVAGET[2], Véronique SEBILLE[1,3]

**Affiliations**

[1] U1246 SPHERE, University of Nantes, University of Tours, INSERM, Nantes, France

[2] CHU Nantes, Department of Addictology and Psychiatry, Nantes, France

[3] CHU Nantes, Department of Methodology and Biostatistics, Nantes, France

[4] INSERM-U1028, CNRS-UMR5292, Lyon Neuroscience Research Center, PSYR[2] Team, University of Lyon, CH Le Vinatier, Lyon, France.

[5] Department of Emergency Psychiatry, Edouard Herriot Hospital, Hospices Civils de Lyon, Lyon, France

**Corresponding author**

Samuel Bulteau, CAPPA Jacques-Prévert, Addiction Medicine and Liaison Psychiatry Department, Hôtel Dieu 3ème Nord, 1 Place Alexis-Ricordeau, 44000 Nantes, France

E-mail: samuel.bulteau@chu-nantes.fr Tel: +33-2.40.08.47.95 Fax: +33-2.40.08.47.98

**Acknowledgments**

STEP (Section for Transcranial Neuromodulation in Psychiatry of the French Association for Biological Psychiatry– AFPBN

The princeps study was supported by the French Ministry of Health, PHRC 2007 (Pr Poulet). The sham venlafaxine was synthesized and delivered by Wyeth (Pfizer) laboratory.

**Conflicts of interest:** The authors declare that they have no conflict of interest

**Abstract**

**Purposes:** Patient-Reported Outcomes are essential to properly assess treatment effectiveness in randomized clinical trial (RCT) for Major Depressive Disorder (MDD). MDD self-assessment may vary over time depending on change in the meaning of patients' self-evaluation of depression, i.e. Response Shift (RS). Our aim was to investigate RS and its impact on different depression domains in a clinical trial comparing rTMS versus Venlafaxine.

**Methods:** The occurrence and type of RS was determined using Structural Equation Modeling applied to change over time in 3 domains (Sad Mood, Performance Impairment, Negative Self-Reference) of the short-form Beck Depression Inventory (BDI-13) in a secondary analysis of a RCT on 170 patients with MDD treated by rTMS, venlafaxine or both.

**Results:** RS was evidenced in the venlafaxine group in the Negative Self-Reference and Sad Mood domains.

**Conclusion:** RS effects differed between treatment arms in self-reported depression domains in patients with MDD. Ignoring RS would have led to a slight underestimation of depression improvement, depending on treatment group. Further investigations of RS and advancing new methods are needed to better inform decision making based on Patient-Reported Outcomes.

**Keywords:** Response Shift-Depression-rTMS-venlafaxine-Patient-Reported Outcomes

## INTRODUCTION

Major depressive disorder (MDD) is a major cause of Quality of Life impairment since it represents the leading cause of years lost to disability (Smith, 2014). Several treatments (pharmacology, psychotherapy, neurostimulation) are available (Patten, 2016), but their respective place and efficacy remains to be refined. Patient-reported

outcome measures (PROMs) are increasingly used to inform healthcare and individual shared decision-making. In depression subjective feeling of recovery is a crucial treatment goal (Zimmerman et al., 2012a; Cohen et al., 2013) and patient's perspective is essential to take into account (Zimmerman et al., 2012b; Tsujimoto et al., 2018; Bulteau et al., 2021; Chevance et al., 2020). Nonetheless, there may be differences in how patients interpret and respond to questions about their health, and therefore how researchers and clinicians may interpret the results in a given context. Some health event or life experience may change patients' standards, priorities, and conceptualizations of a given construct. This may result in measurements not being comparable over time due to changes in patients' internal frame of reference. This change in the meaning of patients' self-evaluation is called response shift (RS) (Sprangers & Schwartz, 1999). It occurs when responses to PROM items do not reflect the construct of interest (e.g. level of depression) in the same way at different points in time (Oort et al., 2009). The definition of RS was recently updated by Vanier et al. (Vanier et al., 2021) where RS is assumed to occur when observed change (e.g. change in observed scores) is not fully explained by target change" (i.e. change in the construct of interest, e.g. depression) as a result of "a change in meaning in self-evaluation of a target construct". Such a discrepancy between observed and target change can occur because of recalibration (a change in one's internal standards such as an acute pain changing the appreciation of a chronic persistent pain), reprioritization (a change in one's values), or reconceptualization related to one's redefinition of the target construct (Vanier et al., 2021).

RS is a *real* phenomenon evidenced in many studies and different pathological contexts such as cancer, stroke, multiple sclerosis, inflammatory bowel disease

(Murata et al., 2020; Ahmed et al., 2005; King-Kallimanis et al., 2010; Lix et al., 2016) and notably in psychiatric diseases (Fokkema et al., 2013; Nolte et al., 2016; Carlier et al., 2019; Smith et al., 2016; Wu, 2016). For instance, Carlier et al., evidenced reconceptualization of mood items (e.g. suicidal ideation became a distinct concept from depression after treatment) and reprioritization (patients placed more value on somatic and cognitive problems after than before treatment) on the Symptom Questionnaire-48 (Carlier et al., 2019). It is likely that some treatments may be more associated with the occurrence of RS than others (e.g. more invasive treatment, or treatment targeting coping skills) and this should be considered in the analyses and their interpretation (Sawatzky et al., 2021). Indeed, if response shift is not taken into account some erroneous conclusion could be drawn. For example, if RS had not been taken into account, psychotherapy could have been found to worsen PROMs scores in an influential clinical trial comparing antidepressants and psychotherapy on BDI scores in a sample of the National Institute Mental Health Treatment of Depression Collaborative Research Program (n=155) (Fokkema et al., 2013). In this case, interpretation and statistical adjustment for RS occurrence allowed the investigators to realize that depression had not actually worsened in the psychotherapy group. The higher levels of depression in this group could be explained by the fact that patients had become better at assessing their level of symptomatology, viewed depression as a more unified concept, and became more aware of their depressive symptoms. (Fokkema et al.,2013). Hence, interpreting and adjusting for RS are essential to obtain unbiased estimates of depression level changes to assess treatments efficacy measured by PROMs over time in MDD.

More generally, ignoring RS may result in failing to meet the health care needs of the population or in inadequate care at a more individual level in terms of over- or under-treatment (Sawatzky et al., 2021).

We need more empirical studies to examine under what circumstances response shift affects which types of interventions in depression. To date, no study has investigated response shift when comparing pharmacotherapy and repetitive Transcranial Magnetic Stimulation (rTMS) in treatment-resistant depression (TRD). In addition, RS may provide insight into differential treatment effectiveness for specific symptoms and/or domains and with regards to reappraisal capacities over time (Bulteau et al., 2019; Verdam et al., 2021). rTMS and medications may not be equal -it is yet not known- in their potential to change attitude towards affect, decrease ruminations and negative self-referencing processes, and enhance cognitive flexibility underlying reappraisal process (Fossati, 2018). It is unknown yet whether medication or brain stimulation could have a differential impact on health state reappraisal during treatment, and especially during the first 4 weeks which corresponds to the expected treatment delay of action at the acute phase. Our hypothesis is that there may exist, some differences between groups and over time in the patients' perception of the level of depression, on different depression domains.

The current study aimed at investigating response shift occurrence, type and magnitude during treatment within a secondary analysis of the largest French multicenter randomized controlled clinical trial for TRD comparing venlafaxine and rTMS (Brunelin et al., 2014).

## METHOD

### *Study design*

Data of a French multicenter randomized double-blind controlled trial comparing three treatment strategies for TRD were used (Brunelin et al., 2014). To evaluate the clinical effect of rTMS or venlafaxine or a combination of both, patients were randomly allocated to a group receiving either: i/ active rTMS combined with active venlafaxine (n = 55), or ii/ active rTMS combined with placebo venlafaxine (n = 60), or iii/ sham rTMS combined with active venlafaxine (n = 55). Stimulation consisted in 1Hz low frequency rTMS (LF-rTMS), 360 pulses/day applied over the right dorsolateral prefrontal cortex at an intensity of 120% of the resting motor threshold. Patients received one session per day for at least 2 days and up to 6 weeks until remission (10 to 30 sessions). Active venlafaxine was initiated 3 days before the first rTMS session (75 mg/day for 3 days) and maintained for 4 weeks from the first day of the rTMS session (150 mg/day). When necessary, based on the clinical judgment of the blind investigator, the dose could be increased to 225mg/day for the next 2 weeks.

*Patient characteristics*

One hundred and seventy patients were randomized in one of the three treatment arms. All participants presented with a single episode or recurrent unipolar non-psychotic major depressive disorder, according to the DSM IV criteria. The participants had to present a score >20 to the Hamilton Depression Rating Scale (HDRS) after the failure of at least one antidepressant treatment delivered at an efficacious dosage for at least 6 weeks. The exclusion criteria were: age under 18, other axis I disorders (except for anxiety disorders) including substance use disorder (except for nicotine), somatic or neurological disorders, failure to respond to

venlafaxine during the current depressive episode, pregnancy, previous rTMS, and rTMS contraindications.

Most of the patients were female (66%, n=111) and were less than 65 years old (84%, n=142). Patients were characterized by treatment resistance (> 2 treatment lines on average), long lasting episode duration (mean=19 months), suicidality (41% attempted suicide), severe functional impairment (49% had long term illness exemption), family history of mood disorders in most cases (59%), and frequent comorbidities especially anxiety (24%) or personality disorders (30%).

*Measures of depression*

This study focused on the depressive symptoms that were self-reported by patients each week with the 13-item short form of Beck Depression Inventory (BDI-13) (Beck & Beamesderfer, 1974). Higher scores on BDI-13 reflect higher levels of depression. The BDI-13 was translated into French (Delay et al., 1963) and its structural validity and concurrent validity were assessed many years ago on small samples of depressed patients (45 and 50 patients, respectively) (Cottraux, 1988; Collet & Cottraux, 1986). No recent studies have comprehensively assessed the validity of the French version of the BDI-13. Hence, the structure of the BDI-13 has been psychometrically investigated in this sample prior to the RS analysis in a previous publication (Bulteau et al., 2021). First, a confirmatory analysis has been performed based on the 3-factor structure of the original version of BDI-13 assessing structural validity and reliability. Fit indices indicated a poor fit and only one dimension was considered as reliable. Hence, an exploratory factor analysis has been performed with orthogonal varimax rotation. An acceptable fit was reached (RMSEA = 0.037, SRMR = 0.064, CFI = 0.954) with a slightly different 3-

factor structure: Sad Mood (including item 1: sadness, item 2: pessimism), Negative Self-Reference (item 3: sense of failure; item 5: guilt; item 6: self-disgust; item 7: suicidal tendencies; item 10: negative self-image), and Performance Impairment (item 4: dissatisfaction; item 8: social withdrawal; item 9: indecision; item 11: working difficulties; item 12: fatigability).

*Statistical analysis*

- A note on structural equation models

All analyses are based on Structural Equation Models (SEM). This relationship is modeled with three parameters: factor loadings, intercepts and residual variances. Applying a SEM analysis consists of specifying a theoretically sounded model with an acceptable fit to the data. A hypothesized model is specified by defining a set of relationships between observed variables such as domain scores and latent variables that represent the inobservable concepts that we want to measure (measurement model) and between the latent variables (structural model). The specification is then considered adequate if the hypothesized model fits well the data. The overall assessment of model fit is based on the $\chi 2$ test statistic for standard maximum-likelihood estimation. Due to the sensitivity of the $\chi 2$ test to sample size, researchers have developed fit indices that are less sensitive to sample size for assessing fit (basically by dividing the $\chi 2$ statistic by its degrees of freedom). The Root mean square error and The comparative fit index are commonly used to assess fit of SEM. The Root mean square error (RMSEA) is a measure of absolute fit and is therefore concerned with the discrepancy due to approximation. It examines the degree of discrepancy between the model and the data through the

observed covariance matrix and is expected to be the lowest possible value (bounded at 0). RMSEA is estimated by the square root of the estimated discrepancy due to approximation per degree of freedom. The comparative fit index (CFI) is a comparative fit index that represents deviations from a null model (a model in which all measured variables are uncorrelated). The CFI analyzes the model fit by examining the discrepancy between the data and the hypothesized model (comparison of chi-square values of the hypothesized model and the null model). CFI ranges from 0 (worst possible fit) to 1 (model fits perfectly).

An important step of SEM is the model modification step. A hypothesized model can be modified to make the model more parsimonious or if it does not have an adequate fit to the data. The model modification is based on the comparison of two nested models, a full model defined by a set of estimated parameters and a restricted model where only a subset of the parameters is estimated. To compare the nested models, a likelihood ratio test (LRT) also named chi-square difference test (Chou 2012) can be performed. The null hypothesis is that the full model does not fit better the data than the restricted model, i.e. the more parsimonious model fits as well as the more complex one. The LRT consists of estimating the difference in chi-square of the two nested models, which follows a chi-square distribution with degrees of freedom calculated as the difference in degrees of freedom of the two models. If the LRT is not significant, it has not been evidenced that the exceeding free parameters of the full model improve the model fit and the restricted model can be retained. If the LRT is significant, the full model is retained.

- Longitudinal model of depression

Before the RS analysis, a longitudinal SEM was performed on the entire study sample to evaluate the fit of a longitudinal model of depression on the data. The scores of Sad Mood, Negative Self-Reference and Performance Impairment were assumed to reflect a unique latent variable representing depression at each time of measurement. The fit of the longitudinal SEM was considered as good (acceptable) if RMSEA<0.05 (<0.08), CFI>0.97 (0.95) (Schermelleh-Engel et al., 2003). The RS analysis was performed only if the fit was acceptable.

- Response shift analysis

Longitudinal SEM were then fitted to the data to detect and quantify RS effects and to estimate longitudinal depression change adjusted, if appropriate, for RS between baseline (randomization) and W4 (4 weeks after randomization), for each domain. SEM assume a linear relationship between the latent variable of depression, and the domain scores of the BDI-13 scale. This relationship is modeled with three parameters: factor loadings, intercepts and residual variances. In case of longitudinal measurement invariance (i.e., no RS), these parameters are invariant over time meaning that change in the observed scores is fully explained by latent change (change in depression). Changes in these parameters are assumed to be indicative of RS. Longitudinal measurement invariance of these parameters was investigated using Oort's procedure based on longitudinal SEM, enabling detecting and quantifying the three types of RS (i.e. recalibration, reprioritization, and reconceptualization) (Oort, 2005). This 4-step procedure operationalizes the three different types of RS as follows: change in patterns of factor loadings (reconceptualization), change in values of factor loadings (reprioritization), intercepts

(uniform recalibration), and residual variances (non-uniform recalibration). Reconceptualization wasn't assessed in this study as it is only possible in multidimensional models.

In step 1, an appropriate measurement model (model 1) is established in which no constraints on parameters related to RS over time are imposed. Namely, the factor loadings, intercepts and residual variances of all BDI-13 domains are non-invariant over time. Good (acceptable) fit is indicated by the following criteria: root mean square error of approximation or RMSEA ≤ 0.05 (0.08), and comparative fit index or CFI ≥ 0.97 (0.95) (Schermelleh-Engel et al., 2003).

In step 2, all RS parameters are constrained to be equal over time constituting a model assuming no RS, i.e., longitudinal measurement invariance on all BDI-13 domains (model 2). The fit of model 2 and of model 1 are compared using a Likelihood Ratio Test (LRT). If the LRT is significant, global occurrence of RS is assumed and the procedure goes on to step 3 to identify the types of RS on the affected BDI-13 domains.

Step 3 consists in a step-by-step improvement of model 2 by relaxing one-by-one RS parameters constraints leading to model 3 accounting for all detected RS. In step 4, the final model (last updated model 3 if the LRT was significant, model 2 otherwise) allows assessing differences in latent variable means over time, adjusted for the detected RS if appropriate, to evaluate longitudinal change in depression.

The Oort's procedure was performed separately in each group to identify occurrence and type of response shift. A longitudinal multi-group SEM including all RS effects previously detected according to group membership was subsequently performed to estimate treatment effects on depression change adjusted for RS. SEM-based effect-size indices of change (Oort, 2005), were estimated in the final multi-group model as standardized response means (Verdam et al., 2017). For each domain affected by RS, the observed change was computed as the difference between observed sample means over time. The contribution of uniform recalibration and reprioritization RS, and of target change (mean change of depression) were computed using the decomposition of change. Non-uniform recalibration does not contribute to observed changes as it has no impact on estimated means.

We assumed no difference in RS parameters between groups at baseline in relation to randomization. All analyses were performed with Stata 16. All data are available upon reasonable request.

To account for missing data, all models were estimated using Maximum Likelihood with Missing Values method (MLMV).

## RESULTS

*Longitudinal model of depression*

A longitudinal SEM performed on the entire study sample indicated a good fit (RMSEA < 0.001, CFI = 1.000) (Figure 1). This structure, named the depression SEM model, was used for RS detection in each group separately.

INSERT Figure 1 about here

*Response shift analyses*

Results of the response shift analysis with Oort procedure in each group are provided in Table 1 reporting fit indices and Likelihood ratio tests (LRT).

INSERT Table 1 about here

- Group 1 (rTMS + venlafaxine) and group 2 (rTMS + venlafaxine placebo)

The same results were obtained in these two groups regarding response shift detection. In step 1, the depression SEM model showed a good fit as indicated in Table 1 (RMSEA < 0.001, CFI = 1.000). The LRT of step 2 was not significant. Hence, no overall RS was assumed in these groups. The final model of step 4 was thus model 2, a model assuming no RS.

- Group 3 (sham rTMS + venlafaxine)

In step 1, the depression SEM model showed a good fit as indicated in Table 1 (RMSEA = 0.026, CFI = 0.998). The LRT of step 2 was significant (p = 0.0103). Hence, model 1 assuming RS was retained. Step 3 was subsequently performed to identify the BDI-13 domains affected by RS and the type of RS. In the first iteration of step 3, model 2 was most improved by relaxing the equality constraint on residual variances of the Sad Mood domain (no-uniform recalibration). In the second iteration, the model was improved by relaxing the equality constraint on intercepts of the Negative Self-Reference domain (uniform recalibration). The third iteration no longer showed any possibility of improving the model, hence stopping the detection of RS. Consequently, the final model of step 4 was accounting for non-uniform recalibration on Sad Mood (changes in residual variances) and uniform recalibration

on Negative Self-Reference (changes in intercepts) to estimate treatment effects on depression change.

- Multi-group model

Estimates from the multi-group SEM model accounting for non-uniform recalibration on Sad Mood and uniform recalibration on Negative Self-Reference in the sham rTMS + venlafaxine group are reported in Figure 2 and Table 2.

INSERT Figure 2 about here

INSERT Table 2 about here

This model showed acceptable fit (RMSEA = 0.054, CFI = 0.971). For the sham rTMS + venlafaxine group, the intercept of Negative Self-Reference increased over time meaning that these patients tended to report higher scores of Negative Self-Reference on average compared to baseline, given similar depression levels over time. Hence, after 4 weeks of treatment, patients receiving sham rTMS + venlafaxine tended to report more intense perceived symptoms within the Negative Self-Reference domain. Furthermore, the residual variance (i.e., variance that is not explained by the latent variable) of Sad Mood decreased over time.

Effects of group and time on depression change, accounting for RS, if appropriate, are presented in Table 3. The following results were similar whether RS was adjusted for or not. For instance, and as expected in a randomized study, adjusted mean estimates of depression showed that all groups started from the same level of depression on average (i.e., test of group effect not significant). The test for time effect was significant in each group showing that depression significantly decreased

in all groups. Moreover, decrease in depression was the same (interaction test with time not significant) in Group 1 (rTMS + venlafaxine) and Group 2 (rTMS + venlafaxine placebo). Contrariwise, the interaction test (Group 3 vs Group 1) was significant and indicated that the level of depression had a larger decrease (time-effect), on average, in Group 3 (sham rTMS+ venlafaxine) than in Group 1 (rTMS and venlafaxine). The following results were however different according to RS adjustment. When RS was accounted for, the interaction test (Group 3 vs Group 2) was significant and indicated that the level of depression had a moderate larger decrease, on average, in Group 3 (sham rTMS+ venlafaxine) than in Group 2 (rTMS + venlafaxine placebo). However, this interaction test was not significant when RS was not accounted for (Table 3).

INSERT Table 3 about here

The estimates of the change in the mean level of depression appear in Figure 3 when RS was accounted for (Figure 3A) or not (Figure 3B). Not adjusting for the response shift occurring only in Group 3 (sham rTMS+ venlafaxine) leaded to underestimate the decrease of depression in this group (adjusted time effect for RS: -1.99, time effect not adjusted for RS: -1.83).

INSERT Figure 3 about here

The observed (reported) change on the Negative Self-Reference domain as well as the RS effects and latent change contributions to the observed change are reported in Table 4 for each treatment group. These estimates are based on the multi-group final SEM model. In the rTMS + venlafaxine and rTMS + placebo venlafaxine groups, observed change equals latent change since no RS was detected. In both of these

groups, the sign of the changes was negative, indicating that depression levels decreased, on average in both groups. In the sham rTMS + venlafaxine group, the RS effect size was 0.27 and the mean latent change was -1.25. Hence, the observed change was equal to -0.98 (i.e., 0.27-1.25). The impact of RS effect, although mild in this study, could lead to underestimating latent change, if not accounted for.

INSERT Table 4 about here

## DISCUSSION

We tested the occurrence of response shift between baseline and after four weeks of treatment in self-reported depressive symptomatology RS (uniform and non-uniform recalibration) was only evidenced in the group of patients who received active venlafaxine combined with sham rTMS. Overall, levels of depression decreased significantly in all groups, regardless of whether RS was accounted for or not. However, ignoring RS in the active venlafaxine + sham rTMS group would have led to a mild underestimation of the improvement of self-reported depressive symptomatology, based on BDI-13. Moreover, decrease in depression according to self-reported BDI-13 could have been wrongly assumed to be similar in sham rTMS + venlafaxine and rTMS + venlafaxine placebo treatment groups. These results draw the attention on RS effects possibly interfering with interpretation of change in PRO when comparing treatments and suggest that RS analysis should be considered before drawing any definite conclusion on difference in treatment efficacy when using self-reported data.

Interestingly, after 4 weeks of treatment, uniform recalibration was detected in the Negative Self-Reference domain (including items such as: feelings of failure, guilt, disappointment, ugliness, auto-aggressive and suicidal thoughts) in the sham rTMS + venlafaxine group only. Uniform recalibration corresponds to the change in the respondent's internal standards of measurement. Specifically, patients tended to report higher ratings on the BDI-13 Negative Self-Reference domain after 4 weeks as compared to baseline, given a similar mean depression level over time. Such a RS effect can impact the interpretation of change in the latent variable (the depression construct we intend to measure). In our study, the improvement of depressive symptomatology would have been underestimated if RS had been ignored. Moreover, it may suggest that some interventions could be more sensitive to RS than others. The decrease in residual variances, i.e., non-uniform recalibration, could reflect less heterogeneity in individual responses on Sad Mood appraisal at W4 as compared to baseline. It may also be that patients with improvement in depression level get better at assessing their levels of depressive symptoms in a more homogeneous way (Fokkema et al., 2013). Of note, non-uniform recalibration does not impact the estimation of change in depression, unlike other types of RS.

We believe that RS should not only be reduced to measurement bias but that it also represents meaningful change. Indeed, RS could also reflect patients' reappraisal of their disease, treatment effects, and coexisting health and life events which may be related to adaptation. Hence, detecting, interpreting and trying to explain RS may provide insight on the way to improve and personalize care. It is therefore of value to

not only adjust for RS when detected but also to see how RS can be interpreted and explained.

Several paths of explanation of an overrating of self-rated symptoms related to negative self-reference processing (i.e. uniform recalibration) in the sham rTMS + venlafaxine group may be put forward. First, after 4 weeks of medication, we could hypothesize that the reduction of psychomotor retardation may have improved patients' insight and consequently influence the way they perceive the items of the questionnaire and their response categories (Silva et al., 2017). Second, explanations of the difference in RS between groups could be related to the Sprangers and Schwartz theoretical model (Sprangers & Schwartz, 1999) which posit that RS may be triggered by *mechanisms* (e.g. coping and emotional regulation strategies such as post-traumatic growth, social comparison) in response to a *catalyst* (a salient health event, e.g. initiation of treatment). *Antecedents* (stable or dispositional characteristics, e.g. symptoms duration, resistance to change, age) can also have direct and indirect effects on potentiating RS. In our study, RS could be due to differences in treatment effects on unmeasured coping abilities or unknown environmental/social events during treatment. We also could hypothesize that reappraisal mechanisms may be impacted differently by rTMS or medications and that there may exist a differential effect of treatments on automatic implicit emotional regulation (early bottom-up effect with amygdala inhibition) and explicit controlled emotional regulation (cognitive reappraisal of one's own negative perception) (Braunstein et al., 2017; Philip et al., 2015; Korb et al., 2011)

*Methodological and clinical perspectives*

As changes in depressive symptomatology can be under or over-estimated if RS is ignored, and some patients following some interventions may be more sensitive to RS than others, we suggest several precautions for the design of comparison studies using PROM in MDD to anticipate and interpret potential RS.

First, we suggest to record some characteristics that may influence the onset of RS or provide some insight about it. In MDD, those variables may be : on the one hand some baseline characteristics less susceptible to change during treatment course (personality, internalized self-image, early stress and traumas, duration of the disease, presence of comorbidities); and on the other hand some variables that may change rapidly over time (symptoms severity, degree of disease acceptance, intensity of self-referencing, engagement in goal-directed behavior, coping skills, cognitive status (especially flexibility within executive functions), insight level, expectations concerning treatment, life priorities and life events).

Secondly, we would recommend researchers to design studies in order to limit bias that could constitute alternative explanations to RS such as: effort justification, social desirability, recall bias. To prevent misunderstanding of results, a social desirability assessment questionnaire can be proposed, or some qualitative research such as cognitive interviews to understand whether the changes in responses to PROMs are attributable to change in meaning of the subjective evaluation, or something else. A second interview can be conducted as the researcher may compare the recalled answers with the previously given answers and invite patients to reflect on the comparison of these responses (Sprangers et al., 2023, *in press*). Eventually, a methodological plan to limit missing data is essential to anticipate in the study design.

Another challenge is to define the adequate time frame for outcomes and RS assessment which is often hypothesized but unknown, notably in MDD. Whether RS is a continuous process or can appear paroxysmally depending on some mechanisms

is still undetermined. It would be interesting in this regard to investigate RS regularly during specific periods without treatment changes and after specific (e.g. cognitive) treatments.

Also, it would be valuable to capture more closely negative self-referencing process at the item-level (emotional regulation, affective self-perception reappraisal, and self-definition) using Rasch Measurement Theory for item-level RS detection which were validated with simulation studies (Blanchin et al., 2020).

Last but not least, RS should also be considered at the individual level. Some methods have already been proposed such as Patient Generated Index (PGI), SeiQol, or semi-structured interviews. Indeed, qualitative methods are the most appropriate methods to reflect the subject reappraisal process at the individual level. We believe that microphenomenology (Petitmengin & Lachaux, 2013) could be a promising approach to define phenomenological markers (we would call « phenomarkers ") useful to stratify patients according to their reappraisal capacities which may be an outcome of biological treatment per se.

*Limitations*

Our sample size of 170 patients can be considered as small for SEM analysis (Wang J.& Wang X., 2012) In the original article describing the procedure on a hypothetical example with two latent variables each measured with four observed variables, Oort (Oort, 2005) indicates that a sample size around 200 subjects would be needed to detect a reprioritization response shift with a medium effect size (power: between 80 and 90%). It is also stated that reconceptualization and recalibration response shifts can be detected with smaller sample sizes. Our

measurement model is less complex with only one latent variable and might therefore require a smaller sample size to detect such response shift effects. Of note, only uniform and non-uniform recalibration response shifts have been detected in our study. We may have missed other RS effects such as reprioritization due to a lack of power. However, many randomized controlled trials in treatment-resistant depression are in this range of number of patients included. Moreover, determining the adequate number of subjects needed for RS analyses remains a difficult task, with regards to the hypotheses to be made (e.g. expected RS effects) and the complexity of the modeling approaches (e.g. multigroup SEM).

Alternative explanations to response shift can also be put forward (Sébille et al., 2021), including misspecification of the SEM model that may alter our ability to detect RS or the interrelations between the different types of RS (e.g. change in residual variances, indicative of non-uniform recalibration, could be the result of change in intercepts, indicative of uniform recalibration, going in different directions) (Oort, 2005; Sébille et al., 2021). In our study, as the final multigroup SEM showed a satisfactory fit, we may trust that this possible alternative explanation may be ruled out.

With regards to the cause or RS an extension of a SEM model can include more explanatory covariates. Verdam et al., showed that anxiety, religion, distress level could be associated with RS in populations presenting with depressive symptoms during psychological interventions (guided internet delivered cognitive behavioural treatment for insomnia, diabetes and meaning-centered group

psychotherapy in cancer survivors) (Verdam et al., 2021). Treatment side-effects could also be taken into account in future studies even if it may constitute a challenge according to variations over time. Of note, incorporating several covariates in SEM usually requires larger sample size.

In addition, performing group-level SEM analyses implies that RS is assumed to be homogeneous within a treatment arm, which is a restrictive and unrealistic assumption. New methods still need to be advanced to explore inter-individual variation in RS (Sébille et al., 2021) and its level of heterogeneity.

## CONCLUSION

In summary, our results strengthen the assumption that RS may be occurring in trials for MDD and differing between treatment arms and self-reported depression domains. Further investigations of RS and new methods are needed to gain more insights into PROs changes over time in MDD.

## Informed Consent

"Informed consent was obtained from all individual participants included in the study."

**Ethical approval:** "All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards."

## References

Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R., 2005. Change in quality of life of people with stroke over time : True change or response shift? *Quality of Life Research, 14*(3), 611627. https://doi.org/10.1007/s11136-004-3708-0

Beck, & Beamesderfer.,1974. Assessment of Depression: The Depression Inventory. Psychological Measurements in Psychopharmacology. *Modern Problems in Pharmacopsychiatry, 7*, 151159.

Blanchin, M., Guilleux, A., Hardouin, J.-B., & Sébille, V., 2020. Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level : A simulation study. *Statistical Methods in Medical Research, 29*(4), 10151029. https://doi.org/10.1177/0962280219884574

Braunstein, L. M., Gross, J. J., & Ochsner, K. N., 2017. Explicit and implicit emotion regulation : A multi-level framework. *Social Cognitive and Affective Neuroscience, 12*(10), 15451557. https://doi.org/10.1093/scan/nsx096

Brunelin, J., Jalenques, I., Trojak, B., Attal, J., Szekely, D., Gay, A., Januel, D., Haffen, E., Schott-Pethelaz, A.-M., Brault, C., STEP Group, & Poulet, E., 2014. The efficacy and safety of low frequency repetitive transcranial magnetic stimulation for treatment-resistant depression : The results from a large multicenter French RCT. *Brain Stimulation, 7*(6), 855863. https://doi.org/10.1016/j.brs.2014.07.040

Bulteau, S., Péré, M., Blanchin, M., Poulet, E., Brunelin, J., Sauvaget, A., & Sébille, V., 2021. Higher Negative Self-Reference Level in Patients With Personality Disorders and Suicide Attempt(s) History During Biological Treatment for Major Depressive Disorder : Clinical Implications. *Frontiers in Psychology, 12*, 631614. https://doi.org/10.3389/fpsyg.2021.631614

Bulteau, S., Sauvaget, A., Vanier, A., Vanelle, J.-M., Poulet, E., Brunelin, J., & Sebille, V., 2019. Depression Reappraisal and Treatment Effect : Will Response Shift Help Improve the Estimation of Treatment Efficacy in Trials for Mood Disorders? *Frontiers in Psychiatry, 10*, 420420.https://doi.org/10.3389/fpsyt.2019.00420

Carlier, I. V. E., van Eeden, W. A., de Jong, K., Giltay, E. J., van Noorden, M. S., van der Feltz-Cornelis, C., Zitman, F. G., Kelderman, H., & van Hemert, A. M., 2019. Testing for response shift in treatment evaluation of change in self-reported psychopathology amongst secondary psychiatric care outpatients. *International Journal of Methods in Psychiatric Research, 28*(3), e1785. https://doi.org/10.1002/mpr.1785

Chevance, A., Ravaud, P., Tomlinson, A., Le Berre, C., Teufer, B., Touboul, S., Fried, E. I., Gartlehner, G., Cipriani, A., & Tran, V. T., 2020. Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals : Qualitative content analysis of a large international online survey. *Lancet Psychiatry, 7*(8), 692702.

Chou, C.-P., & Huh, J. 2012. Model modification in Structural Equation Modeling. In *Handbook of Structural Equation Modeling* (1st edition., pp. 232–246). New-York, NY: Guilford Press.

Cohen, R. M., Greenberg, J. M., & IsHak, W. W., 2013. Incorporating multidimensional patient-reported outcomes of symptom severity, functioning, and quality of life in the Individual Burden of Illness Index for Depression to measure treatment impact and recovery in MDD. *JAMA Psychiatry, 70*(3), 343350. https://doi.org/10.1001/jamapsychiatry.2013.286

Collet, L., & Cottraux, J.,1986. [The shortened Beck depression inventory (13 items). Study of the concurrent validity with the Hamilton scale and Widlöcher's retardation scale]. *L'Encephale, 12*(2), 7779.

Cottraux,J., 1988. Depressive Cognitions of obsessive-compulsive patients : A factorial analysis of the shorter form of the Beck depression inventory. In *Cognitive Therapy* : An update (Popuu Press). C. Perris & M. Eisemann, eds.

Delay, J., Pichot, P., Lemperiere, T., & Mirouze, R.,1963. [THE NOSOLOGY OF DEPRESSIVE STATES. RELATION BETWEEN ETIOLOGY AND SEMEIOLOGY. 2. RESULTS OF BECK'S QUESTIONAIRE]. *L'Encephale, 52*, 497504.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P., 2013. Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment, 25*(2), 520531. https://doi.org/10.1037/a0031669

Fossati, P., 2018. Is major depression a cognitive disorder? *Revue de Neurologie, 174*(4), 212215. https://doi.org/10.1016/j.neurol.2018.01.365

Gao, W., Chen, S., Biswal, B., Lei, X., & Yuan, J., 2018. Temporal dynamics of spontaneous default-mode network activity mediate the association between reappraisal and depression. *Social Cognitive and Affective Neuroscience, 13*(12), 12351247. https://doi.org/10.1093/scan/nsy092

King-Kallimanis, B. L., Oort, F. J., & Garst, G. J. A., 2010. Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *Advances in Statistical Analysis, 94*(2), 139156. https://doi.org/10.1007/s10182-010-0129-y

Korb, A. S., Hunter, A. M., Cook, I. A., & Leuchter, A. F., 2011. Rostral anterior cingulate cortex activity and early symptom improvement during treatment for major depressive disorder. *Psychiatry Research, 192*(3), 188194. https://doi.org/10.1016/j.pscychresns.2010.12.007

Lix, L. M., Chan, E. K. H., Sawatzky, R., Sajobi, T. T., Liu, J., Hopman, W., & Mayo, N., 2016. Response shift and disease activity in inflammatory bowel disease. *Quality of Life Research, 25*(7), 17511760. https://doi.org/10.1007/s11136-015-1188-z

Murata, T., Suzukamo, Y., Shiroiwa, T., Taira, N., Shimozuma, K., Ohashi, Y., & Mukai, H., 2020. Response Shift-Adjusted Treatment Effect on Health-Related Quality of Life in a Randomized Controlled Trial of Taxane Versus S-1 for Metastatic Breast Cancer : Structural Equation Modeling. *Value Health, 23*(6), 768774. https://doi.org/10.1016/j.jval.2020.02.003

Nolte, S., Mierke, A., Fischer, H. F., & Rose, M., 2016. On the validity of measuring change over time in routine clinical assessment: A close examination of item-level response shifts in psychosomatic inpatients. *Quality of Life Research, 25*(6), 13391347. https://doi.org/10.1007/s11136-015-1123-3

Oort, F. J., 2005. Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587598.

Oort, F. J., Visser, M. R., & Sprangers, M. A. 2009. Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. Journal of clinical epidemiology, 62(11), 1126–1137.

Patten, S. B., 2016. Updated CANMAT Guidelines for Treatment of Major Depressive Disorder. *Canadian Journal of Psychiatry, 61*(9), 504505. https://doi.org/10.1177/0706743716660034

Petitmengin, C., & Lachaux, J.-P., 2013. Microcognitive science : Bridging experiential and neuronal microdynamics. *Frontiers in Human Neuroscience, 7*, 617. https://doi.org/10.3389/fnhum.2013.00617

Philip, N. S., Carpenter, S. L., Ridout, S. J., Sanchez, G., Albright, S. E., Tyrka, A. R., Price, L. H., & Carpenter, L. L., 2015. 5Hz Repetitive transcranial magnetic stimulation to left prefrontal cortex for major depression. *Journal of Affective Disorders, 186*, 1317.

Sawatzky, R., Kwon, J.Y., Barclay, R., Chauhan, C., Frank, L., van den Hout, W.B., Nielsen, L.K., Nolte, S., Sprangers, M.A.G.; Response Shift – in Sync Working Group., 2021. Implications of response shift for micro-, meso-, and macro-level healthcare decision-making using results of patient-reported outcome measures. *Quality of Life Research, 30*(12):3343-3357. doi: 10.1007/s11136-021-02766-9.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H., 2003. Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit. *Measures Methods of Psychological Research,* 2374.

Sébille, V., Lix, L. M., Ayilara, O. F., Sajobi, T. T., Janssens, A. C. J. W., Sawatzky, R., Sprangers, M. A. G., Verdam, M. G. E., & Response Shift – in Sync Working Group., 2021. Critical examination of current response shift methods and proposal for advancing new methods. *Quality of Life Research 30*(12):3325-3342

Silva, R. de A. da, Mograbi, D. C., Camelo, E. V. M., Santana, C. M. T., Landeira-Fernandez, J., & Cheniaux, E., 2017. Clinical correlates of loss of insight in bipolar depression. *Trends in Psychiatry and Psychotherapy, 39*(4), 264269. https://doi.org/10.1590/2237-6089-2017-0007

Smith, D., Woodman, R., Harvey, P., & Battersby, M., 2016. Self-Perceived Distress and Impairment in Problem Gamblers : A Study of Pre- to Post-treatment Measurement Invariance. *Journal of Gambling Studies, 32*(4), 10651078. https://doi.org/10.1007/s10899-016-9598-6

Smith, K., 2014. Mental health : A world of depression. *Nature, 515*(7526), 181. https://doi.org/10.1038/515180a

Sprangers, M. A., & Schwartz, C. E., 1999. Integrating response shift into health-related quality of life research : A theoretical model. *Social Science & Medicine, 48*(11), 15071515. https://doi.org/10.1016/s0277-9536(99)00045-3

Sprangers, M.A., Sawatzky, R., Vanier, A., Böhnke, Jan R., Sajobi, T., Mayo, N., Lix, L., Verdam M., Oort, F., Sébille, V., and the Response Shift – in Sync Working Group, 2023. Implications of the syntheses on definition, theory and methods conducted by the Response Shift – in Sync Working Group. *Quality of Life Research*, in press.

Sawatzky, R., Kwon, J. Y., Barclay, R., Chauhan, C., Frank, L., van den Hout, W. B., Nielsen, L. K., Nolte, S., Sprangers, M. A. G., & Response Shift – in Sync Working Group, 2021. Implications of response shift for micro-, meso-, and macro-level healthcare decision-making using results of patient-reported outcome measures. *Quality of Life Research*, *30*(12), 3343–3357.

Tsujimoto, E., Tsujii, N., Mikawa, W., Ono, H., & Shirakawa, O. , 2018. Discrepancies between self- and observer-rated depression severities in patients with major depressive disorder associated with frequent emotion-oriented coping responses and hopelessness. *Neuropsychiatric disease and treatment*, *14*, 2331–2336. https://doi.org/10.2147/NDT.S175973

Vanier, A., Oort, F. J., McClimans, L., Ow, N., Gulek, B. G., Böhnke, J. R., Sprangers, M., Sébille, V., Mayo, N., & Response Shift - in Sync Working Group, 2021. Response shift in patient-reported outcomes: definition, theory, and a revised model. *Quality of Life Research, 30*(12), 3309–3322. https://doi.org/10.1007/s11136-021-02846-w

Verdam, M. G., Oort, F. J., & Sprangers, M. A.,2016. Using structural equation modeling to detect response shifts and true change in discrete variables: an application to the items of the SF-36. *Quality of Life Research*, *25*(6), 1361–1383. https://doi.org/10.1007/s11136-015-1195-0

Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G., 2017. Structural equation modeling-based effect-size indices were used to evaluate and interpret the impact of response shift effects. *Journal of Clinical Epidemiology, 85*, 3744. https://doi.org/10.1016/j.jclinepi.2017.02.012

Verdam, M. G. E., van Ballegooijen, W., Holtmaat, C. J. M., Knoop, H., Lancee, J., Oort, F. J., Riper, H., van Straten, A., Verdonck-de Leeuw, I. M., de Wit, M., van der Zweerde, T., & Sprangers, M. a. G., 2021. Re-evaluating randomized clinical trials of psychological interventions : Impact of response shift on the interpretation of trial results. *PloS One, 16*(5), e0252035. https://doi.org/10.1371/journal.pone.0252035

Wu, P.-C., 2016. Response Shifts in Depression Intervention for Early Adolescents. *Journal of Clinical Psychology, 72*(7), 663675. https://doi.org/10.1002/jclp.22291

Zimmerman, M., Martinez, J., Attiullah, N., Friedman, M., Toba, C., & Boerescu, D. A., 2012a. Why do some depressed outpatients who are not in remission according to the hamilton depression rating scale nonetheless consider themselves to be in remission? *Depression and Anxiety, 29*(10), 891895. https://doi.org/10.1002/da.21987

Zimmerman, M., Martinez, J., Attiullah, N., Friedman, M., Toba, C., & Boerescu, D. A., 2012b. Symptom differences between depressed outpatients who are in remission according to the Hamilton Depression Rating Scale who do and do not consider themselves to be in remission. *Journal of Affective Disorders, 142*(13), 7781. https://doi.org/10.1016/j.jad.2012.03.044

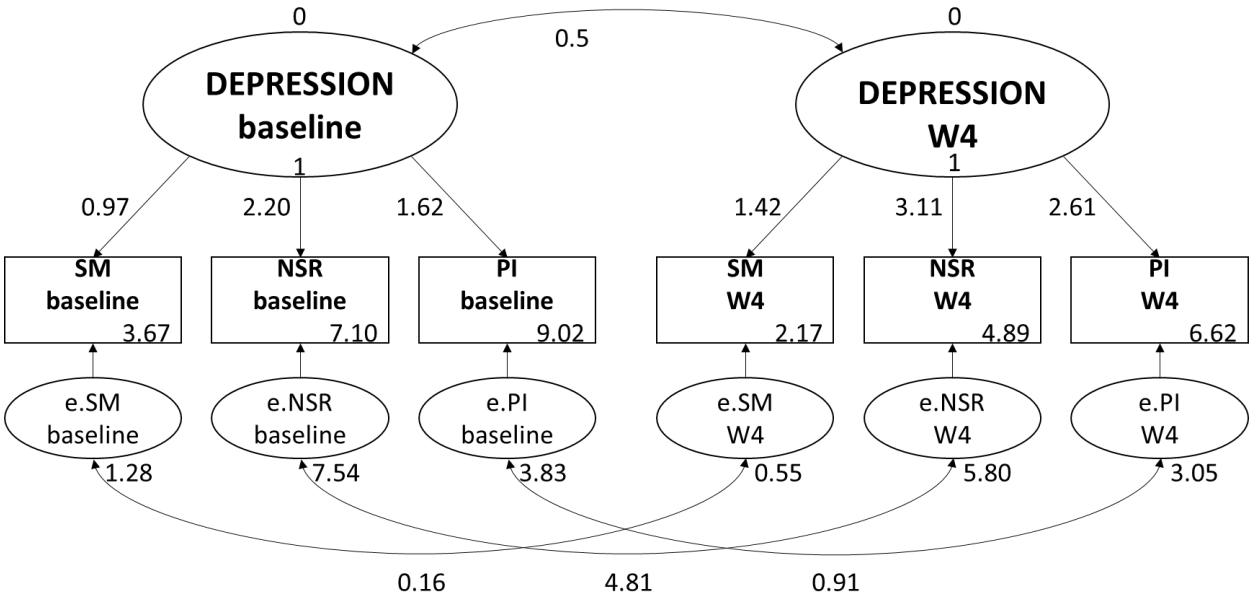



Figure 1: Depression model. Longitudinal SEM between randomization (baseline) and 4 weeks after (W4) performed on the entire study sample.

Fit indices: $\chi^2$=4.66, df=5, p<0.0001, RMSEA<0.0001, CFI=1.000

Circles represent latent variables of depression, squares represent observed variables (domain scores of BDI-13). Double-sided arrows represent correlations between variables. Constraints: latent variable means=0, latent variable variances=1

SM: Sad Mood, NSR: Negative Self- Reference, PI: Performance Impairment

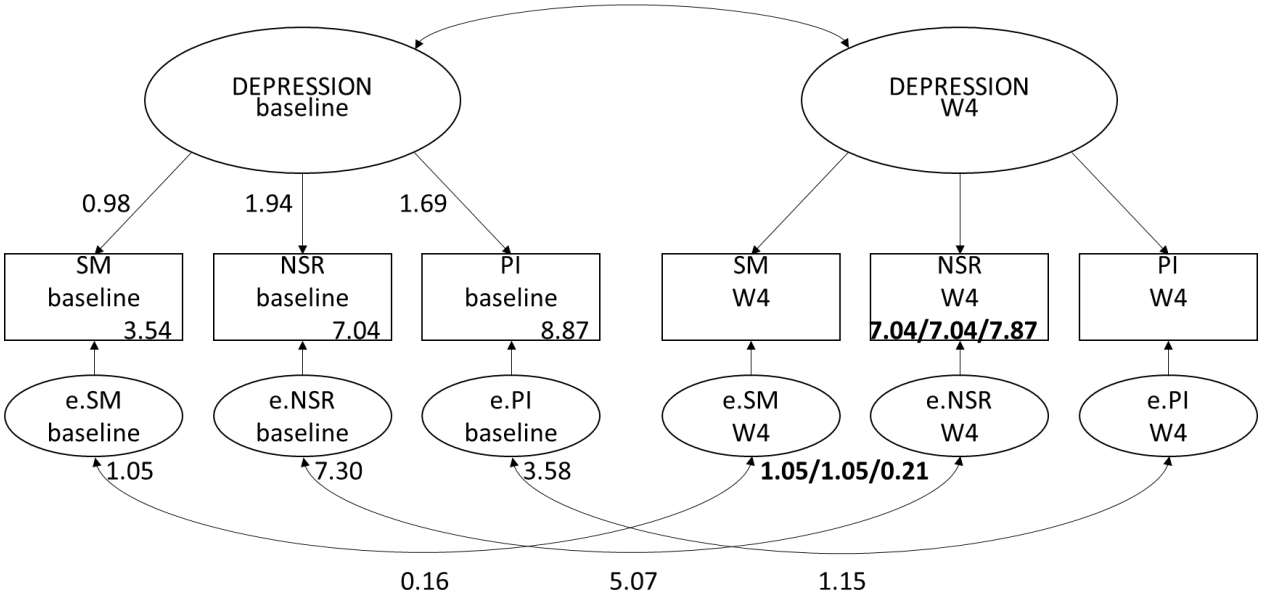
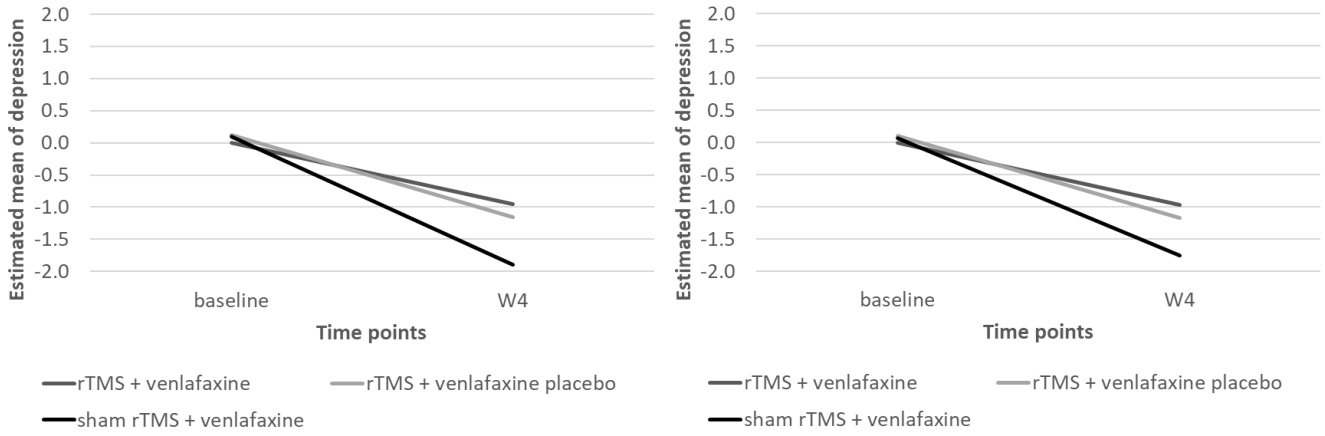



Figure 2: Longitudinal multi-group SEM of the response shift analysis between randomization (baseline) and 4 weeks after (W4). Estimates at week 4 are represented for each group respectively if they have changed over time (rTMS + venlafaxine, rTMS + venlafaxine placebo, sham rTMS + venlafaxine) separated by a slash. Estimations of means, variances and covariances of latent variables are omitted.

Fit indices: $\chi^2$=62.903, ddf=54, p=0.190, RMSEA=0.054 [0.000,0.104], AIC=3575.089, CFI=0.971, TLI=0.976

Circles represent latent variables of depression, squares represent observed variables (domain scores of BDI-13). Double-sided arrows represent correlations between variables.

SM: Sad Mood, NSR: Negative Self- Reference, PI: Performance Impairment




A: Response shift accounted for

B: No response shift assumed

Figure 3: Latent variable (depression) mean change over time by treatment group
baseline: randomization, W4: 4 weeks after randomization

| | Group 1 rTMS + venlafaxine (N=54) | | | | | Group 2 rTMS + venlafaxine placebo (N=58) | | | | | Group 3 sham rTMS + venlafaxine (N=54) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\chi^2$ | df | p value | RMSEA | CFI | $\chi^2$ | df | p value | RMSEA | CFI | $\chi^2$ | df | p value | RMSEA | CFI |
| Model 1 | 1.86 | 5 | 0.868 | 0 | 1 | 4.09 | 5 | 0.536 | 0 | 1 | 5.188 | 5 | 0.393 | 0.026 | 0.998 |
| Model 2 | **14.14** | **12** | **0.292** | **0.058** | **0.982** | **15.21** | **12** | **0.230** | **0.067** | **0.963** | 23.60 | 12 | 0.023 | 0.133 | 0.887 |
| LRT | stat | df | p | | | stat | df | p value | | | stat | df | p value | | |
| model 2 vs model 1 | 12.28 | 7 | 0.092 | | | 11.12 | 7 | 0.133 | | | 18.41 | 7 | 0.010 | | |
| Models 3 | $\chi^2$ | df | p value | RMSEA | CFI | $\chi^2$ | df | p value | RMSEA | CFI | $\chi^2$ | df | p value | RMSEA | CFI |
| NURC Sad Mood | NA | | | | | NA | | | | | 16.14 | 11 | 0.136 | 0.092 | 0.950 |
| URC Negative Self-Reference | NA | | | | | NA | | | | | **11.46** | **10** | **0.323** | **0.052** | **0.986** |

Table 1: Fit indices for each group and each step of response shift analysis. Likelihood ratio tests (LRT) are also reported. The results for final models are in bold. Criteria of good (acceptable) fit: RMSEA≤0.05 (0.08) and CFI≥0.97 (0.95)
model 1: model without any constraints on response shift parameters
model 2: model assuming no RS, i.e., longitudinal measurement invariance on all BDI-13 domains
NURC: non-uniform recalibration, URC: uniform recalibration, NA: not applicable, RMSEA: root mean square error, CFI: comparative fit index
stat: test statistic, df: degrees of freedom

| | | Group 1 rTMS + venlafaxine (N=54) | | Group 2 rTMS + venlafaxine placebo (N=60) | | Group 3 sham rTMS + venlafaxine (N=55) | |
|---|---|---|---|---|---|---|---|
| | | *baseline* | *W4* | *baseline* | *W4* | *baseline* | *W4* |
| | | est. (s.e.) | est. (s.e.) | est. (s.e.) | est. (s.e.) | est. (s.e.) | est. (s.e.) |
| FACTOR LOADINGS | SM | 0.98 (0.15) | 0.98 (0.15) | 0.98 (0.15) | 0.98 (0.15) | 0.98 (0.15) | 0.98 (0.15) |
| | NSR | 1.94 (0.30) | 1.94 (0.30) | 1.94 (0.30) | 1.94 (0.30) | 1.94 (0.30) | 1.94 (0.30) |
| | PI | 1.69 (0.24) | 1.69 (0.24) | 1.69 (0.24) | 1.69 (0.24) | 1.69 (0.24) | 1.69 (0.24) |
| INTERCEPTS | SM | 3.54 (0.17) | 3.54 (0.17) | 3.54 (0.17) | 3.54 (0.17) | 3.54 (0.17) | 3.54 (0.17) |
| | NSR | 7.04 (0.37) | 7.04 (0.37) | 7.04 (0.37) | 7.04 (0.37) | 7.04 (0.37) | **7.87 (0.58)** |
| | PI | 8.87 (0.30) | 8.87 (0.30) | 8.87 (0.30) | 8.87 (0.30) | 8.87 (0.30) | 8.87 (0.30) |
| RESIDUAL VARIANCES | SM | 1.05 (0.16) | 1.05 (0.16) | 1.05 (0.16) | 1.05 (0.16) | 1.05 (0.16) | **0.21 (0.18)** |
| | NSR | 7.30 (0.89) | 7.30 (0.89) | 7.30 (0.89) | 7.30 (0.89) | 7.30 (0.89) | 7.30 (0.89) |
| | PI | 3.58 (0.48) | 3.58 (0.48) | 3.58 (0.48) | 3.58 (0.48) | 3.58 (0.48) | 3.58 (0.48) |
| DEPRESSION | | | | | | | |
| mean | | 0 (constrained) | -0.96 (0.32) | 0.12 (0.23) | -1.15 (0.32) | 0.10 (0.24) | -1.89 (0.39) |
| variance | | 1 (constrained) | 2.44 (0.93) | 1.09 (0.42) | 1.93 (0.78) | 1.11 (0.43) | 1.74 (0.68) |
| | | *baseline, W4* | | *baseline, W4* | | *baseline, W4* | |
| covariance | | 0.42 (0.34) | | 0.91 (0.44) | | 0.75 0.36) | |
| | | SM | NSR | PI | | | |
| Residual covariances (baseline, W4) | | 0.16 (0.14) | 5.07 (0.91) | 1.15 (0.48) | | | |

Table 2: Estimations and standard errors of the parameters of the longitudinal multi-group SEM of the response shift analysis between randomization (baseline) and 4 weeks after (W4). Estimates of response shift related parameters (factor loadings, intercepts and residual variances) at week 4 are in bold if they have significantly changed over time.

SM: Sad Mood, NSR: Negative Self- Reference, PI: Performance Impairment

| | Response shift accounted for | | | | No response shift assumed | | | |
|---|---|---|---|---|---|---|---|---|
| | **estimate** | **s.e.** | $\chi^2$ | **pvalue** | **estimate** | **s.e.** | $\chi^2$ | **pvalue** |
| **Time effect** | | | | | | | | |
| Group 1 (rTMS + venla-faxine) | -0.96 | 0.32 | 8.94 | 0.003 | -0.97 | 0.32 | 9.33 | 0.002 |
| Group 2 (rTMS + venla-faxine placebo) | -1.27 | 0.27 | 21.71 | <0.001 | -1.28 | 0.27 | 22.32 | <0.001 |
| Group 3 (sham rTMS + venlafaxine) | -1.99 | 0.37 | 28.95 | <0.001 | -1.83 | 0.35 | 27.28 | <0.001 |
| **Overall group effect (J0)** | | | 0.30 | 0.863 | | | 0.21 | 0.901 |
| **Time*group interaction** | | | | | | | | |
| Group 2 vs Group 1 | -0.32 | 0.36 | 0.78 | 0.376 | -0.31 | 0.36 | 0.74 | 0.391 |
| Group 3 vs Group 1 | -1.03 | 0.41 | 6.46 | 0.011 | -0.85 | 0.39 | 4.70 | 0.030 |
| Group 3 vs Group 2 | -0.72 | 0.33 | 4.61 | 0.032 | -0.55 | 0.32 | 2.87 | 0.091 |
| **Response shift effects** | | | | | | | | |
| NURC Sad Mood | -0.84 | 0.20 | 17.25 | <0.001 | | | | |
| URC Negative Self-Reference | 0.83 | 0.49 | 2.84 | 0.092 | | | | |

Table 3: Multi-group SEM model accounting for response shift (Non-uniform recalibration on Sad Mood and uniform recalibration on Negative Self-Reference in the sham rTMS + venlafaxine group) or not.
Time effect: depression mean difference over time for a given group
Group effect: depression mean difference across groups at a given time
NURC: non-uniform recalibration, URC: uniform recalibration

| Negative Self-Reference | Group 1<br>rTMS + venlafaxine<br>N=54 | Group 2<br>rTMS + placebo venlafaxine<br>N=58 | Group 3<br>sham rTMS + venlafaxine<br>N=54 |
|---|---|---|---|
| Observed change<br>(reported change) | -0.49 | -0.83 | -0.98 |
| RS contribution<br>(uniform recalibration) | 0.00 | 0.00 | 0.27 |
| Mean latent change<br>(mean change of depression) | -0.49 | -0.83 | -1.25 |

*Table 4. Estimated effect-size indices of change on Negative Self-Reference domain. Model estimation based on the whole sample (final multi-group SEM model including RS detected by Oort's procedure applied to each group)*