Measurement of symptom change following web based psychotherapy – statistical characteristics and analytical methods for measuring and interpreting change

1E. Karin, 1,3B.F, Dear, 2G. Z. Heller, 1M. Gandy & 1,3N. Titov

1 eCentreClinic, Department of Psychology, Macquarie University.
2 Department of Statistics, Faculty of Science and Engineering, Macquarie University.
3 MindSpot Clinic, Department of Psychology, Macquarie University.
Highlights

- In this paper we explore the statistical characteristics of depressive symptom change (PHQ9; Kroenke, Spitzer, & Williams, 2001) associated with a large web-based psychotherapy sample (ICBT; n=1096), and compare between two common ways to measure and interpret symptom change (linear and proportional change).

- Results demonstrated that within both treatment and waitlist conditions, symptoms changed proportionally to baseline (e.g. 50-55% treatment related change across individuals from different baselines, and ±30% change for individuals the waitlist). Additional features such as (1) a strong relationship between an individual’s baseline score and the rate of symptom change, and (2) positive distributional skewness at post treatment, also suggest proportional change.

- Analyses demonstrate the measurement of proportional symptom change also enabled (1) a more accurate in predicting symptom outcomes (measurement error reduction of over 40%), (2) more interpretable estimate of change (percentage improvement) which is aligned with the aims of treatment (remission of symptoms), (3) a clearer ability to differentiate between treatment related and non-treatment related symptom change, and (4) overcome an artificial increase between the estimation of treatment efficacy and baseline symptom severity.

- This research suggests that the measurement and interpretation of symptom change as a proportional function (percentage improvement) can be more suitable than common linear alternatives such as Cohen’s d. Although some statistical jargon is used, the intention of this paper is to convey the statistical methodology, and the take home messages about symptom measurement in a way that is approachable to any practitioner.
Abstract

Objective: The aims of this study were to (1) explore the underlying characteristics of depressive symptom change (PHQ9) following psychotherapy, and (2) compare the suitability of different ways to measure and interpret symptom change. A treatment sample of web-based psychotherapy participants (n=1098), and a waitlist sample (n=96) were used to (1) explore the statistical characteristics of depressive symptom change, and (2) compare the suitability of two common types of change functions; linear and proportional change.

Methods: These objectives were explored using hypotheses which tested (1) the relationship between baseline symptoms and the rate of change, (2) the shape of symptom score distribution following treatment, and (3) measurement error associated with linear and proportional measurement models.

Result: Findings demonstrated that: (1) individuals with severe depressive baseline symptoms reduced by greater symptom scores than individuals with mild baseline symptoms (11.4 vs. 3.7). However, as a percentage measurement, change remained similar across individuals with mild, moderate or severe baseline symptoms (50-55%); (2) positive skewness was observed in PHQ9 score distributions following treatment; (3) models that measured symptom change as a proportional function resulted in greater model-fit, and reduced measurement error (< 30%).

Conclusions: This study suggests that symptom scales, sharing an implicit feature of score bounding, are associated with a proportional function of change. Selecting statistics that overlook this proportional change (e.g., Cohen’s d) is problematic and leads to (1) artificially increased estimates of change with higher baseline symptoms, (2) increased measurement error, and (3) confounded estimates of treatment efficacy and clinical change.
Introduction

Accurate measurement of treatment related change is a key part of psychotherapy research (Kroenke, Monahan & Kean, 2015; Spring, 2007; Wise, 2004) and the investigation of treatment efficacy (Flay, Biglan, Boruch, Castro, Gottfredson, Kellam, ... & Ji, 2005; Gottfredson, Cook, Gardner, Gorman-Smith, Howe, Sandler & Zafft, 2015; Laurenceau, Hayes & Feldman, 2007). For example, measurable change in symptoms of anxiety and depression is often used as the primary means to research and test the safety of emerging treatments (Titov, Dear, Staples, Bennett-Levy, Klein, Rapee, ... & Purtell, 2015), describe the clinical trajectory of participants in treatment (Gunn, Elliott, Densley, Middleton, Ambresin, Dowrick, ... & Griffiths, 2013), the cost-effectiveness of treatment (Sobocki, Ekman, Ågren, Runeson, & Jönsson, 2006), and the comparison of treatments (Gyani, Shafran, Layard & Clark, 2011). For this reason, the ability to measure change with accurate and valid methods is critical for psychotherapy (Altman & Simera, 2016; Laurenceau, et al., 2007).

To increase the validity and accuracy of change measurement in psychotherapy, several statistical and clinical methods are employed. The most common methodology in psychotherapy research is the combined use of standardized scales, such as standardized symptom scales of anxiety (Choi, Schalet, Cook & Cella, 2014) or depression (Kroenke et al., 2015; Schalet, Cook, Choi & Cella, 2014), and the use of statistical analyses, such as Cohen’s d effect sizes, that measure and interpret the rate of change in treatment (Flay, et al., 2005; Gottfredson, et al., 2015; Laurenceau, et al., 2007). It is important to acknowledge that many
types of standardized scales are available for measuring and interpreting change in treatment (e.g. clinical interviews, measurement of behaviour or quality of life; Snyder, Aaronson, Choucair, Elliott, Greenhalgh, Halyard, ... & Santana, 2012), and that change can be statistically estimated through various statistical methods (e.g. Baldwin, Fellingham & Baldwin, 2016; Keller, 2003). However, from the wide range of possible methods for measuring treatment outcomes (Clarke, 2007), the use of standardized scales, primarily symptom scales, in combination with effect sizes, and primarily Cohen’s $d$, are the most influential. For example, symptom scales and effect sizes are used to evaluate treatment related change and treatment efficacy within psychotherapy trials (Horn, & Gassaway, 2007; Lakens, 2013; Schulz, Altman & Moher, 2010), epidemiological studies (Bower, Kontopantelis, Sutton, Kendrick, Richards, Gilbody, ... , Liu, 2013; Clark, 2011), meta-analytic studies of various treatments (Newby, McKinnon, Kuyken, Gilbody & Dalgleish, 2015), and are even mandated within clinical guidelines for reporting in clinical trials such as CONSORT (Schulz et al., 2010), TREND (Des Jarlais, Lyles, & Crepaz, 2004), STROBE (Von Elm, Altman, Egger, Pocock, Gøtzsche, Vandenbroucke, & Strobe Initiative, 2014) and others (Altman & Simera, 2016).

Notwithstanding the common use of both symptoms scales and effects sizes for measuring psychotherapeutic related change, little research is currently available to verify or refute the use of different statistical methods for measuring and interpreting symptom change (Hiller, Schindler & Lambert, 2012; McMillian, Gilbody & Richards, 2010). For example, the use of effect sizes, such as Cohen’s $d$, is based on statistical assumptions that change is linear. In technical terms, by employing effect sizes, researchers assume that the symptom change that follows treatment is average, constant and representative of the average change experienced by any participating individual (Ellis, 2010; Lakens, 2013). Put another way, if an average individual with moderate depressive symptoms prior to treatment (e.g., PHQ-9
scores of 10-15) would improve by five points on a symptom scale, an individual with severe baseline symptoms (e.g., PHQ-9 score of 20 to 27) would be expected to demonstrate the same rate of improvement (e.g. five points). Similarly, under the linear assumption, a group of participants with different baseline symptoms (e.g. mild, moderate or severe baseline symptoms), undertaking the same therapy, would be expected to have similar effect sizes between groups (e.g. 1.0). However, in contrast to the common use of statistics that assume change is linear, there are two lines of research to suggest that real world symptoms change may occur as a proportional function from baseline. First, psychological treatment studies often describe an increased rate of clinical change within samples of increased baseline symptom severity (Bower, et al., 2013; Boettcher, Hasselrot, Sund, Andersson & Carlbring, 2014; Hedman, Lindefors, Andersson, Andersson, Lekander, Ruck & Ljotsson, 2013). Second, common symptom scales, such as the Patient Health Questionnaire-9 Item scale (PHQ-9) (Kroenke, Spitzer, & Williams, 2001), the Generalised Anxiety Disorder 7-Item scale (GAD-7) (Kroenke, Spitzer, Williams, Monahan & Löwe, 2007) and prominent others (e.g. K10)(Kessler, Andrews, Colpe, Hiripi, Mroczek, Normand, ... , & Zaslavsky, 2002), often demonstrate an implicit design feature of score bounding at minimal symptoms. This bounding within symptoms scales should theoretically imply that, under effective treatment, all individuals would reduce their symptoms down to the same endpoint of minimal levels (Kroenke et al., 2015; Sobocki et al., 2006), and that the rate of change would systematically depend on an individual’s symptom at baseline (Driessen & Hollon, 2010; Thase, Simons, Cahalane, McGeary & Harden, 1991) .

From a statistical point of view, identifying the characteristics of symptom change, and employing a suitable statistical analysis which captures the underlying function of change, can fundamentally impact both the measurement and interpretation of clinical outcomes (Baldwin et al., 2016; Fitzmaurice & Laird, 2012; Hoffman, 2015; Liang and
For example, under circumstances where change is proportional in nature, the selection of a proportional statistical analysis can greatly increase the accuracy and validity of estimating longitudinal clinical change (Fitzmaurice & Laird, 2012; Liang and Zenger, 1986), the detection of moderators of symptom change (Castellani, Rajaram, Gunn, & Griffiths, 2016), the classification of subgroups, such as remitters or non-responders (Panagiotakopoulos, Lyras, Livaditis, Sgarbas, Anastassopoulos & Lymberopoulos, 2010), as well as the ability to research other objectives (Pocock, Clayton & Stone, 2015). For this reason, the function of symptom change must be researched, and more clearly understood. Such research could verify, refute, and draw out the implication for using well-established statistical methods (e.g. effect sizes, linear statistics), and emerging alternatives (percentage improvement, generalised linear statistics) for measuring and interpreting change in treatment. In addition, researching the function and characteristics of symptom change has the potential to inform researchers and the broader community about the type of change individuals in treatment are likely to experience.

The present study

This study aims to (1) explore the fundamental statistical characteristics of treatment related depressive symptom change; and (2) compare the implications from measuring and interpreting clinical change through effect sizes, such as Cohen’s $d$, against emerging alternatives such as percentage improvement (proportional, generalised longitudinal linear statistics) (Hiller et al., 2012; McMillian et al., 2010).

This study employed a large sample of individuals ($n = 1098$) who underwent web based psychotherapy (internet-delivered cognitive behaviour therapy; ICBT; Titov, Dear, Staples, Terides, Karin, Sheehan, …, & McEvoy, 2015) for symptoms of depression (PHQ-9; Kroenke et al., 2001). Although web based psychotherapy represents a distinct type of psychotherapy, the use of web based treatments, which standardizes treatment materials and participant
engagement through automatisation, can be seen as an opportunity for researching symptom change with high internal validity and minimum methodological interference.

The statistical characteristics of symptom change were explored with three steps. Initially, the relationship between baseline symptoms and the rate of change was explored. In line with previous clinical studies that suggest that more severely symptomatic participants demonstrate increased effect sizes (Bower et al., 2013; Driessen, Cuijpers, Hollon & Dekker, 2010) it was hypothesised that individuals with increased symptom at baseline would also demonstrate increased rates of symptom change (H1). Second, the shape of symptom score distribution prior and following treatment were explored. In line with the suggestion that symptoms scores are bounded at minimal symptoms (Kroenke et al., 2001; Kroenke et al., 2007), the distributions of pre-treatment and post-treatment depression symptom levels were hypothesized to show evidence of positive skewness and kurtosis, at both pre-treatment and post treatment (H2). Third, the measurement error associated with linear and proportional measurement models was compared. In line with the characterisation of symptom change as proportional it was hypothesized that those statistical methods that measure symptoms change as a proportional function would be associated with reduced measurement error and indicate great statistical fit to real symptom data in treatment (H3). Finally, an additional effort was taken to explore the patterns of depressive symptom change within a control group (n=96). This addition was designed to explore the pattern of symptom change that is not specific to treatment.

**Methods**

**The sample**

The present study combined clinical data from three published randomised control trials (RCTs), all of which evaluated internet-delivered cognitive behaviour therapy interventions...
ICBT) for symptoms of depression and anxiety (Dear, Staples, Terides, Karin, Johnston, Gandy,… & Titov, 2015; Titov, Dear, Johnston, Lorian, Zou, Wootton, & Rapee, 2013; Titov, N., Dear, B. F, Staples, L., Terides M.D., Karin, E., Sheehan J, … & McEvoy, 2016). These interventions were almost identical in structure and therapeutic content. All trials were delivered using the same evidence based online treatment approach (Macquarie University Online Model; Titov, et al., 2015), were conducted within the same research clinic, the eCentreClinic (www.ecentreclinic.org), and resulted in similar symptom reductions. Together, these trials represent a large random intake of adults into treatment over period of two years with a total of 1262 adult participants, of whom 1098 (87%) were successfully assessed at both pre and post treatment time points.

In order to be included in these trials, participants were selected on the basis of (1) demonstrating at least mild symptoms of depression or anxiety (a minimum score ≥ 5 on either the Participant Health Questionnaire 9-item; PHQ-9; or the Generalised Anxiety Disorder Scale 7-item; GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006); (2) being over the age of 18 and under the age of 65; (3) being an Australian resident; and (4) having internet access for the period of the trial. In addition, applicants who reported a score of 3 (considered severe) on item 9 of the PHQ-9 measuring suicidal risk, were referred to another service.

Additional demographic and symptom characteristics are shown in Table 1. For both the treatment and waitlist control conditions.

<table>
<thead>
<tr>
<th>Table 1 – Sample Demographics</th>
<th>Collated treatment sample (n=1098)</th>
<th>Control sample (n=96)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender proportions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>30% (330)</td>
<td>53% (51)</td>
</tr>
<tr>
<td>Age (Mean, SD) in years</td>
<td>56.3 (13.0)</td>
<td></td>
</tr>
<tr>
<td>Sample proportions using medication during the course</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>32% (351)</td>
<td>53% (51)</td>
</tr>
<tr>
<td>Sample proportions with Marital status</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
married/de facto
65% (713) 47% (45)
single/never married
10% (109) 22% (21)
separated/divorced/widowed
25% (274) 31% (30)

**Sample proportions with employment**

| Employed | 58% (636) | 51% (49) |

**Sample proportions with differing Education levels**

| High school | 16% (176) | 41% (39) |
| Vocational | 28% (307) | 25% (24) |
| Degree | 56% (615) | 35% (37) |

**Pre-treatment symptom levels**

| PHQ9 prior to treatment (Mean, SD) | 11.73 (4.8) | 10.9 (4.7) |
| GAD7 prior to treatment (Mean, SD) | 10.9 (4.5) | 9.5 (4.5) |

PHQ9; patient health questionnaire nine item; GAD7 Generalised anxiety disorder seven item scale

**Symptom Measure**

The PHQ-9 was employed as the primary outcome variable, measuring the presence and severity of depressive symptoms (Kroenke, et al., 2001). The PHQ-9 is widely used in clinical trials (Clark, 2011; Titov et al., 2015), comprising of 9 items, with high internal consistency, and high sensitivity to the presence and change of clinical depression diagnoses (Kroenke, et al., 2001). Scores on the PHQ-9 correspond to the cumulative experience of common depressive symptoms, over the preceding 2-week period. Cumulative scores range from 0 to 27, and scores are clinically interpreted as falling within 5 categories: no depression symptoms (total score: 0 to 4), mild depression symptoms (total score: 5 to 9), moderate depression symptoms (total score: 10 to 14), moderately severe depression symptoms (total score: 15 to 19) and very severe depression symptoms (total scores: 20 to 27). Symptom scores were modified with a small constant added (0.001), to ensure that plausible values of zero symptoms at post-treatment were represented in the model when statistically modelling proportional functions, such as logarithmic link functions.
Analytical plan

The function of symptom change was explored with three separate steps, corresponding to the three hypotheses.

Hypothesis 1: relationship between baseline symptoms and the rate of change

The first hypothesis was tested by examining the relationship between baseline symptoms and the rate of symptom change. To evaluate the relationships between baseline and symptom change, the rate of symptom change was examined within the five subgroups of individuals of different baseline PHQ9 score bands (e.g., minimal to no symptoms to very severe depression symptoms). Within each subgroup, the rate of change was approximated with generalised estimated equation models (GEE) (Liang & Zeger, 1986), multi-level models (Fitzmaurice & Laird, 2012) and raw means. These methods represent common longitudinal statistical methods in clinical trials (Hubbard, Ahern, Fleischer, Van der Laan, Lippman, Jewell & Satariano, 2010) and the estimation of change through all three GEE, mixed models and raw scores intended to clarify that the underlying function of symptom change could be identified when using various statistical models.

Under a linear pattern of symptom change, participants of any baseline symptoms would be expected to show a similar rate of improvement overall. That is, an average symptom change score that would be observed across individuals, irrespective of the severity of their symptoms at baseline (Lakens, 2013). In contrast, under a proportional pattern of symptom change, participants presenting with increased baseline symptom severity would likely show larger symptom change compared to those individuals with mild or moderate baseline symptoms (Baldwin, et al., 2016).
Hypothesis 2: The Distribution of Symptom Scores

To test the second hypothesis, the distributions of depression symptoms scores at both pre-treatment and post-treatment were evaluated for evidence of skewness. In this step, if the dataset would present with statistically normal distribution of symptom scores at both time points, the symptoms change over time would be considered as linear. In contrast, if symptoms changed as a proportional function from baseline, positive skewness should be observed, particularly at post treatment, where individuals from various baseline symptoms would shift and concentrate around the symptom score band of minimal symptoms. Graphical and numerical explorations of pre-post score distributions were included.

Hypothesis 3: Linear and Multiplicative model fit

To test the third hypothesis, the relative measurement accuracy of models that represent either linear or proportional symptom change were compared. Specifically, this step compared model fit statistics and the remaining unexplained (residual) variance associated with each function of change. Both mixed models and GEE models were run initially as models that assume change was linear, represented through models that specified a normal scale of the dependent variable and an identify link function. Following this, alternative statistical models were compared; which specified a gamma scale and a log link function, representing models that assumed change was proportional. Generally, the gamma scale is considered a suitable method for data showing signs of skewness and multiplicative change function (Baldwin et al., 2016), however, the selection of the gamma scale does not imply that alternative multiplicative statistical methods (e.g. negative binomial scale, Poisson scale, or zero inflated models) would be less effective.

The formulas below emphasizes the difference in statistical notation between the multiplicative model, and linear models.
Multiplicative model

\[ Y_{ij} \sim \text{Gamma} (\mu_{ij}, \alpha) \]

\[ \log (\mu_{ij}) = \beta_0 + \beta_{ij} \]

Linear additive model

\[ Y_{ij} = \beta_0 + \beta_{ij} + \text{error variance}_{ij} \]

\[ \text{error variance}_{ij} \sim \text{N} (0, 1) \]

\[ i = 1, \ldots, 1096; \; j = 0, 1 \]

\[ t_j = \{ 0 \text{ (time }= \text{ pre-treatment); } 1 \text{ (time }= \text{ post treatment)} \]  

\[ \beta_0 \text{ is the random intercept at pre-treatment and } \beta_{ij} \text{ is the treatment effect of change over time} \]

With more formal statistical notation, the multiplicative effect within the log link model is created when the intercept, \( \beta_0 \), or baseline symptoms, is multiplied by the treatment effect \( \beta_{ij} \), the estimate of exponential change following treatment. Specifically,

\[ \hat{\mu}_{\text{baseline}} = e^{\hat{\beta}_0} \]

\[ \hat{\mu}_{\text{posttreatment}} = e^{\hat{\beta}_0} \times e^{\hat{\beta}_{t1}} \]

\[ \hat{\mu}_{\text{posttreatment}} = \hat{\mu}_{\text{baseline}} \times e^{\hat{\beta}_{t1}} \]

where \( \hat{\mu}_{\text{baseline}} \times e^{\hat{\beta}_{t1}} \) represents change from baseline as a multiplicative function where

\[ \hat{\mu}_{\text{baseline}} + \beta_{tj} \] represents the measurement of the additive function where change is an average quantity added to baseline scores.
The suitability of either model types were evaluated through model fit statistics. generated using SAS 9.4 software (SAS Institute, Cary NC).

Specifically, the Quasilikelihood under the Independence model Criterion, QIC statistic (Pan, 2001) for GEE models, and Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for Mixed effects models (Akaike, 1973), compared between linear (additive) and generalised linear (proportional) models. Within all AIC, BIC and QIC model fit estimates, relatively lower scores imply overall reduced variance, and overall increase measurement accuracy.

In addition to model fit statistics, the measurement error associated with the assumption that symptom change was either a fixed average score, or a percentage improvement scores, was compared. In this step, measurement error was created for each participant by comparing the predicted post treatment score under each change assumption (e.g. 5 PHQ9 points change, or 50% from baseline), against a known participant outcome score at post-treatment. The difference between the expected symptom outcome and actual treatment outcome effectively represent measurement error under the two change assumptions, akin to residual scores and measurement error variance. The pattern of residuals created under either assumption of symptom change was explored in two ways. First, the total quantity of error variance under each function was compared. Second, measurement residuals were graphically explored under each function of symptom change, by comparing the increase or decrease of residuals for individuals with different baseline symptom scores.

**Results**

**Hypothesis 1: relationship between baseline symptoms and the rate of change**

In the first step, the relationship between baseline symptom severity and the quantity of symptom change was explored graphically. Figure 1, illustrating PHQ9 change as a linear
function, and Figure 2, illustrating PHQ9 change as a proportional change from baseline, both demonstrate the symptom change on the Y axis, within each of the PHQ9 baseline symptom bands (X axis). In addition, the symptom change observed within the waitlist condition is included as a dotted trend line, illustrating the trend of nonspecific change in symptoms within each bands of symptom severity at baseline.

**[Figure 1]**

Figures 1 illustrates an increased rate of symptom change which associated closely with increased baseline symptoms. In Figure 1, individuals with severe baseline symptoms were observed to reduce by as much as three-fold when compared to individuals with mild baseline symptoms (11.4 vs 3.7 respectively). In addition, participants with severe symptoms in the control group demonstrated a sizable reduction in symptoms even when treatment was not applied. This non-specific symptom related change was pronounced to the extent that individuals with severe baseline symptoms in the control group demonstrated higher symptom reductions than individuals with moderate symptoms in treatment (7 points vs 6 points respectively). That is, as a linear effect, the nonspecific symptom change within the control condition was larger than the treatment related symptom change of individuals with moderate symptoms.

**[Figure 2]**

Figure 2 illustrates the proportional, percentage change of symptoms within each of the mild, moderate, moderately severe and severe subgroups. The figure illustrates that as a proportional change, an average treatment related change of 50-55% was observed across all
subgroups of individuals who started with at least mild symptoms at baseline. Of note, the rate of proportional improvement in treatment (50%-55%) was greater than the nonspecific change experienced by individuals with severe baseline symptoms in the waitlist conditions (35%). That is, the measurement of change as a percentage change resulted in a clearer differentiation of treatment specific and non-specific change.

Table 2 includes the numerical descriptions of change, for the both the treatment and control conditions. Table 2 also includes effect sizes that were calculated within the treatment group as a whole, and the effect size demonstrated by individuals in the mild, moderate, moderately severe and severe bands of baselines symptoms. Notably in Table 2, individuals with mild depressive symptoms show smaller effects (1.59) when compared to patients with more severe symptoms (3.9).

<table>
<thead>
<tr>
<th>Function of change</th>
<th>Initial symptom severity category</th>
<th>Observed means and standard deviations</th>
<th>Total sample change estimate (GEE) †</th>
<th>Change estimate (GEE) †</th>
<th>Control group change (nonspecific effect) †</th>
<th>Effect sizes (Cohen's d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear function</td>
<td>Minimal (n=72)</td>
<td>Pre-treatment Mean=2.8, SD=1.2</td>
<td>0.61 [0.1, 1.178]</td>
<td>-2 [-3.7, -0.32]</td>
<td>0.32 [0.01, 0.63]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score range 0-4</td>
<td>Post Mean=2.2, SD=2.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mild (n=345)</td>
<td>Pre-treatment Mean=7.4, SD=1.3</td>
<td>3.7 [3.3, 3.4]</td>
<td>-0.1 [-1, 0.8]</td>
<td>1.59 [1.43, 1.74]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score range 5-9</td>
<td>Post Mean=3.7, SD=3.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate (n=381)</td>
<td>Pre-treatment Mean=12.0, SD=1.4</td>
<td>6 [5.7, 6.3]</td>
<td>6.2 [5.8, 6.6]</td>
<td>0.3 [-0.5, 1.1]</td>
<td>2.34 [2.19, 2.49]</td>
</tr>
<tr>
<td></td>
<td>Score range 10-14</td>
<td>Post Mean=5.8, SD=3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderately severe (n=244)</td>
<td>Pre-treatment Mean=16.7, SD=1.4</td>
<td>8.7 [8.9, 3.8]</td>
<td>0.5 [-1.0, 1.1]</td>
<td>2.54 [2.33, 2.74]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score range 15-19</td>
<td>Post Mean=8.1, SD=5.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Severe (n=56)</td>
<td>Pre-treatment Mean=20.9, SD=0.86</td>
<td>11.4 [10.1, 12.7]</td>
<td>7.3 [4.6 to 10.1]</td>
<td>3.9 [3.45, 4.36]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score range 20-27</td>
<td>Post Mean=9.4, SD=5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportional function</td>
<td>Minimal (n=72)</td>
<td>Pre-treatment Mean=2.8, SD=1.2</td>
<td>52% [50%, 52%]</td>
<td>21% [-1 to 39]</td>
<td>-61% [-78 to -44]</td>
<td></td>
</tr>
</tbody>
</table>
### Hypothesis 2: The Distribution of Symptom Scores

Figures 3 illustrates the distribution of PHQ-9 symptom scores, both prior and following treatment. These histograms illustrated a slight positive skewness of scores at pre-treatment, with fewer individuals presenting within the severely symptomatic band as compared to the mild and moderate bands. In contrast, at post-treatment, increasing positive skewness was observed, where most individuals who reduced their symptoms become concentrated within the mild to minimal symptom ranges.

[Figure 3]

The numerical estimates of the skewness are collated in Table 3.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Time point</th>
<th>Skewness statistics (with standard error)</th>
<th>Average baseline symptoms (with SD)</th>
<th>Effect sizes (Cohen's d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment sample (n=1098)</td>
<td>Pre</td>
<td>† <strong>0.271 (0.071)</strong></td>
<td>12.3 (4.5)</td>
<td>1.31 [1.25,1.37]</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>† <strong>1.359 (0.076)</strong></td>
<td>5.8 (4.4)</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Pre</td>
<td>0.178 (0.109)</td>
<td>10.9 (4.9)</td>
<td>-0.01 [-0.17,0.19]</td>
</tr>
</tbody>
</table>
** Estimation of statistical significantly above and beyond zero at 95% confidence level.
† Statistical significance beyond 0.05 alpha, on a Shaprio Wilk test for distribution normality; significance is indicative that normal distribution is not supported within the observed sample

Taken together, both numerically and graphically, the distributions of symptoms scores demonstrated significant positive skewness that increased at post treatment.

**Hypothesis 3: Linear and Multiplicative model fit**

Table 4 collated the goodness of fit statistics taken from models that specified either a proportional or linear function of change.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Method of change specified</th>
<th>QIC (GEE model)</th>
<th>AIC (Mixed)</th>
<th>BIC (Mixed)</th>
<th>Total variance (PHQ9 $\sigma^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment sample (n=1098)</td>
<td>Linear (normal scale)</td>
<td>52457.6</td>
<td>14059.8</td>
<td>14071.3</td>
<td>16.716</td>
</tr>
<tr>
<td></td>
<td>Proportional (Gamma scale)</td>
<td>2020.5</td>
<td>4041.8</td>
<td>4053.3</td>
<td>24.122</td>
</tr>
</tbody>
</table>

Model fit criterion of generalised Chi-Square derived from SAS software, version 9.3. QIC, AIC and BIC model fit statistics derived from SPSS version 22. ††† Confidence intervals based on the multiplicative longitudinal GEE model specified in the analytical plan.

In Table 4, models that specified a proportional function of symptom change demonstrated a several-fold improvement in the model fit statistics, within both the GEE and mixed models; including reduced QIC statistics, reduced AIC and reduced BIC estimates. Table 4 also collated the measurement error associated with the prediction that change occurred as a linear change of six points, or as a percentage improvement (52% reduction from baseline). A notable reduction in the total estimate of PHQ9 error variance was evident when a proportional function of change was assumed ($\sigma^2$ of 16.716 vs. 24.122). This result
demonstrated that by characterising change as a proportional function, the measurement error and remaining unknown individual variation reduced by over 30%.

The measurement error associated with either assumptions that change was linear (six points) or proportional (52%) were graphically explored. Figure 4 illustrates the residual error (y axis) across individuals who started treatment with different baseline symptoms (x axis). In the figure, individuals with mild and severe baseline symptoms can be observed to substantially underestimate, or overestimate the rate of symptom change when linear change (six points) is predicted. In contrast, when change was predicted to be proportional (52%), baseline symptoms no longer associated with the rate measurement error. Further, under the proportional assumption, the predicted symptom outcome could be accurately predicted within a single point across individuals with different baselines (marked with dots horizontal lines). In contrast, under the linear assumption, the prediction of symptom outcome become systematically erroneous with baseline severity (a range of up to sixteen points between mild and severe).

[Figure 4]

Discussion

This study aimed to investigate the statistical characteristic of symptom change in treatment, and compare between different ways to measure and interpret symptom change. Using a web based psychotherapy sample (n = 1098), as well as a waitlist control condition (n=96), the statistical characterisation of depressive symptom change (PHQ9) was explored in three steps, corresponding to three proposed hypotheses.

Testing of the first hypothesis demonstrated support for the characterisation of symptom change as a proportional function (H1), through a clear association between the symptoms severity at baseline and the rate of change. In contrast, as a proportional estimate of change,
individuals in treatment demonstrated a consistent rate of proportional symptom change within all subgroups with mild, moderate, moderately severe and severe baseline symptom (50-55%). Critically, the dependency between symptom change and baseline symptom severity was also observed in the waitlist condition; with mild and severe participants changing proportionally in their symptoms even when treatment was not applied. Testing of the second and third hypotheses also illustrated support for the characterisation of symptom change as proportional function, with symptom score distributions presenting with positive skewness, particularly following treatment (H2). Similarly, increased model fit, and reduced measurement error was observed when the treatment sample was statistically modelled with an underlying proportional function of change (H3).

The analyses within this study are novel in that they characterise the function of depressive symptom change, and compare between different statistical methods for measuring and interpreting the symptom change within treatment as well as non-treatment conditions. The findings suggest that common psychotherapy symptom scales (e.g., PHQ-9) are impacted by a feature of natural bounding at minimal symptoms, which is the suspected culprit for the resulting (1) non-normal distributions at post treatment, (2) the dependency between baseline symptoms and rate of change, and (3) the improved model fit for techniques that assume longitudinal change is proportional to baseline.

The current findings raise two potentially critical implications for the ability to measure and interpret psychotherapy change in combination with symptoms scales. First, the inappropriate use of linear statistics, such as Cohen’s $d$, when change is proportional would lead to artificially higher estimates of clinical efficacy, both in treatment and in control conditions. For example, in the current study, individuals with severe baseline symptoms demonstrated effect sizes that increased by nearly threefold (3.9) when compared to individuals with mild symptoms (1.59); even when the same treatment was applied. This is
problematic as linear estimates of change such as Cohen’s $d$ are strongly associated with baseline severity and not with quality or the effectiveness of treatment. This finding is broadly consistent with the data within previous psychotherapy studies showing increased effect sizes with samples of increased symptoms, even when similar treatments are applied (Bower, et al., 2013; Driessen, et al., 2010; Kroenke et al., 2001).

Second, the current findings support a well-established statistical idea posing that the selection of a statistical analysis must match the characteristics of the dataset in order to arrive at valid and accurate statistical measurement, interpretation and conclusions (Flay et al., 2005; Field & Wilcox, 2017). In this context of depressive symptoms scales, the use of proportional statistical analyses resulted in (1) improved statistical modelling of treatment effects, (2) an improved ability to determine what a treatment effect is (50-55%), and what a non-treatment effect is (35%), as well as for (3) establishing a clinical effect that is robust across individuals with various baseline symptoms (50-55%). The measurement and interpretation of change as proportional improvement from baseline can also be concretely and easily interpreted as an estimate of change (e.g. percentage improvement). Further, in the context of treatment, percentage improvement and percentage change estimates seem to reflect the ideal of treatment (reducing symptoms to minimal)(Kroenke et al., 2015; Sobocki et al., 2006). For these reasons, measuring and interpreting change as a fundamentally proportional function can hold critical implications for clinical research that is reliant on accurate and interpretable measurement. For example, researchers seeking to identify clinical moderators, compare between treatments, estimate cost effectiveness or classify individual effects, are likely to be positively impacted with a suitable choice of analytics that capture the underlying statistical function of change (Castellani et al., 2016; Panagiotakopoulos et al., 2010).
Although the measurement and interpretation of symptom change as a proportional change show promise to increase the accuracy and interpretability of clinical change, several statistical and clinical limitations should be considered about the results of this study. Primarily, the results of this study should be considered as (1) preliminary, (2) specific to a symptom scale of depressive symptoms (PHQ-9), and (3) specific to one kind of treatment model (the Macquarie University online model). To address these limitations, statistical replication must be conducted across different symptom scales, and treatment models. In this way, the characterisation of symptom remission as a proportional pattern can be verified and considered as a broader psychotherapy measurement principle.

Further, it is important to consider that measurement and interpretation of symptom change as a proportional function is at odds with the widely accepted use of linear statistics in psychotherapy. From one point of view, linear statistics, such as Cohen’s $d$, are successful as an established measurement standard that can be used to compare change estimates between trials and across clinical instruments (Spring, 2007). This use of effect sizes has resulted in both enormous amounts of aggregated evidence about the effects of psychotherapy (Newby et al., 2015), and for this reason, it is understandable clinical researchers would continue to use this standard for measuring and interpreting symptom change. However, should symptom change occur as a proportional function, the measurement and interpretation of treatment related change would substantially improve by matching appropriate statistical analysis to the characteristics of the function of symptom change (Baldwin et al., 2016; Field & Wilcox, 2017; Verkuilen & Smithson, 2012). A possible solution to this dilemma would be to report both the effect size and percentage estimates of change side by side. In this way, the change that occurs in treatment can be more accurately reported, evaluated and compared between trials.
Finally, this study does not weigh whether the change rate of 50-55% could be evaluated as the same treatment related effect across individuals with severe or mild baseline symptoms. For example, a symptom reduction demonstrated by individuals with severe baseline symptoms could be interpreted as a more substantive clinical effect than an equivalent symptom reduction achieved with individuals with mild or moderate symptoms (Judd, Schettler, Rush, Coryell, Fiedorowicz & Solomon, 2016). To address these limitations, additional research into the experience of individuals in treatment could determine whether individuals with different baseline symptoms consider the proportional remission pattern an equally satisfactory treatment outcome. For example, Zimmerman and colleagues (Zimmerman, McGlinchey, Posternak, Friedman, Attiullah & Boerescu, 2006) consider the measurement of patient functionality, positive mental health and optimism alongside the reduction in depressive symptoms. These additional measures, could verify and elaborate on the experience of individuals in treatment within various symptom bands, shedding more light on the universality, or segmentation of the 50-55% improvement effect.

In summary, this study aimed to explore the underlying pattern of symptom change, and compare between different methods for measuring and interpreting the depressive symptom change that follows treatment (web based psychotherapy). This study has combined evidence of increased rate of change with increased baseline symptoms (H1), score distributions that become increasingly skewed following treatment, (H2) and increased measurement accuracy achieved by statistical methods that assume change is proportional (H3), to suggest that the fundamental function of symptom change is proportional. The promise of matching these characteristics of proportional symptom change to a suitable statistical analysis is important for all (1) statistical modelling and the prediction of treatment effects, (2) an improved ability to differentiate treatment and non-specific symptom change, as well as for (3) determining an estimate of treatment related change that will not sway with increased baseline symptoms.
Replication of these preliminary findings are essential, within additional depressive symptom scales, other types of psychological conditions, and across different treatments modalities.

References


Horn, S. D., & Gassaway, J. (2007). Practice-based evidence study design for comparative effectiveness research. Medical care, 45(10), S50-S57. doi: 10.1097/MLR.0b013e318070c07b


Kroenke, K., Monahan, P. O., & Kean, J. (2015). Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. Journal of clinical epidemiology, 68(9), 1085-1092. doi.org/10.1016/j.jclinepi.2015.03.023


for anxiety and depression. Psychiatric Services, 66(10), 1043-1050. doi.org/10.1176/appi.ps.201400477


