Can we trust clinical significance when evaluating clinical outcomes? An investigation of the reliability, validity, flexibility, and stability of clinical significance methodology

Fiona R Ronk, BSc (Hons)

School of Psychology
The University of Western Australia

Year of submission: 2014

This thesis is presented for the degree of Doctor of Philosophy at
The University of Western Australia
Abstract

Measuring patient change during mental health treatment is vital in both evaluating the effectiveness of treatment programs and in guiding clinical decisions about a patient’s future care. Clinical significance methodology (Jacobson, Follette, & Revenstorf, 1984) provides a way of quantitatively conceptualising changes patients make between two time periods, usually pre-treatment and post-treatment. This method takes into consideration both (a) whether a patient has made a change that is considered statistically reliable, and (b) whether a patient resembles a member of the functional, healthy population, or the dysfunctional, treatment-seeking population (Jacobson & Truax, 1991). The potential benefits of using the methodology are becoming more well-known and reports of the clinical significance of research findings are being increasingly recommended by psychology journals. However, can we trust its conclusions and interpretations? Several assumptions have been made regarding the use of clinical significance methodology, many of which have been largely untested, and clear recommendations for its use are lacking. Therefore, further exploration is required before clinical significance methodology can be confidently relied upon to yield accurate evaluations of treatment outcome that lead to valid interpretations. This thesis aims to contribute to the field of mental health outcome evaluation by describing and exploring assumptions made regarding clinical significance methodology and making recommendations for its future use.

Five studies were conducted using data from inpatient samples. Firstly, it was found that clinical significance classification rates are largely consistent when different methods are used to calculate them. Secondly, classifications of recovered demonstrate ecological validity. That is, patients who are considered recovered have scores on other variables that align with the concept of recovery. Thirdly, patients find change that
places them into the ‘functional’ range of a measure to be more meaningful than change that is large in magnitude. Next, it was demonstrated that clinical significance methodology can be flexed to take into account different types of patient change, and a new category of *recovering* was proposed. Finally, it was demonstrated that recovery should be considered to be neither solely a process nor solely an outcome, but rather a dynamic process that is different for each individual. The findings of this thesis support the continued use of clinical significance methodology, and recommendations are made for future use and reporting of clinical significance findings. Given the small but significant proportion of patients who make no change or deteriorate during treatment, it is deemed necessary for future research to firstly determine the validity of these undesirable outcomes and then examine factors associated with these, with the ultimate goal of improving treatment outcomes for patients.
Table of Contents

Abstract ................................................................................................................................................. 1

Manuscripts Arising from this Thesis ................................................................................................. 5

Author Contributions ......................................................................................................................... 7

Acknowledgments ............................................................................................................................... 9

Section One: General Introduction

Chapter One ............................................................................................................................................... 13

The importance of evaluating outcomes .......................................................................................... 15

Methods of outcome evaluation ......................................................................................................... 15

What is clinically significant change? ................................................................................................. 17

Jacobson-Truax method for evaluating clinical significance ............................................................... 19

An Assumption that the Approach Taken to Assess Clinical Significance Does Not Matter ................. 22

An Assumption that Clinical Significance Classifications are Equally Valid ....................................... 26

An Assumption that Clinical Significance Evaluations are Equally Appropriate for Individuals at Every Severity Level ........................................................................................................ 29

An Assumption that Clinical Significance Methodology Considers Classifications as Outcomes Rather than Processes ........................................................................................................... 32

Thesis Aims and Outline ...................................................................................................................... 34

Section Two: Reliability and validity of clinical significance classifications

Chapter Two .............................................................................................................................................. 47

How consistent are clinical significance classifications when calculation methods and outcome measures differ? ..................................................................................................................... 49

Chapter Three ....................................................................................................................................... 79

Foreword ................................................................................................................................................. 81
Comparing the validity of clinically significant change classifications yielded by the Jacobson-Truax method and the Hageman-Arrindell method..................87

Chapter Four ...............................................................................................................123
Which component of calculating clinically significant change more validly describes post-treatment recovery? .................................................................125

Section Three: Exploring the versatility of clinical significance methodology

Chapter Five ..................................................................................................................147
Foreword .....................................................................................................................149
Assessing clinical significance of treatment outcomes using the DASS-21....155

Chapter Six ..................................................................................................................185
Investigating the stability of clinically significant mental health outcomes.....187

Section Four: General Discussion

General Discussion ....................................................................................................203

Findings and recommendations.................................................................................206

How reliable are clinical significance classifications when clinically significant change is measured differently? .................................................................206

Are classifications of recovery valid?........................................................................208

Can clinical significance classifications flex to consider different patient goals?.................................................................................................................211

Are clinical significance categories stable following treatment?.................213

Future research........................................................................................................215

General conclusions..............................................................................................216
Manuscripts Arising from this Thesis

Chapter 2


Chapter 3


Chapter 4

**Ronk, F. R.,** Hooke, G. R., & Page, A. C. (in preparation). Which component of calculating clinically significant change more validly describes post-treatment recovery?

Chapter 5


Chapter 6

Author Contributions

Chapter 2: Fiona Ronk (70%) and Andrew Page (20%) designed the study; Geoff Hooke (10%) managed the data collection; Fiona Ronk conducted the statistical analyses and wrote the first draft of the manuscript. All authors contributed to and have approved of the final manuscript.

Chapter 3: Fiona Ronk (70%) and Andrew Page (20%) designed the study; Geoff Hooke (10%) and Fiona Ronk contributed to data collection. Fiona Ronk conducted the statistical analyses and wrote the first draft of the manuscript. All authors contributed to and have approved of the final manuscript.

Chapter 4: Fiona Ronk (70%) and Andrew Page (20%) designed the study; Geoff Hooke (10%) and Fiona Ronk contributed to data collection. Fiona Ronk conducted the statistical analyses and wrote the first draft of the manuscript. All authors contributed to and have approved of the final manuscript.

Chapter 5: Fiona Ronk (60%) and Andrew Page (20%) designed the study; James Korman (10%) and Geoff Hooke (10%) managed the data collection. Fiona Ronk conducted the statistical analyses and wrote the first draft of the manuscript. All authors contributed to and have approved of the final manuscript.

Chapter 6: Fiona Ronk (70%) and Andrew Page (20%) designed the study; Geoff Hooke (10%) and Fiona Ronk contributed to data collection. Fiona Ronk conducted statistical analyses and wrote the first draft of the manuscript. All authors contributed to and have approved of the final manuscript.
Acknowledgments

I have been fortunate to have received the support of so many people while writing this PhD thesis.

Firstly, my most sincere gratitude to Winthrop Professor Andrew Page, for his constant encouragement and guidance. I have appreciated his brilliant sense of humour, especially when the work was particularly challenging. Andrew’s passion for applying research to clinical practice has inspired me, and I know I will continue to embody the scientist-practitioner model throughout my future research endeavours and in my clinical practice.

This research could not have been possible without the assistance of Moira Munro, the Clinical Improvement Team, and all the clinical staff at Perth Clinic. I am honoured to have worked with such an effective team, who are truly passionate about improving outcomes for their patients.

In particular, I would like to thank Geoff Hooke, whose practical assistance in implementing this research has been invaluable. His professionalism, enthusiasm, and quick wit has made this process so enjoyable.

Thank you to James Korman and his team at the Summit Medical Group in New Jersey for their enthusiasm for this research and their data contribution.

I would like to thank the many patients who have contributed to this research.

This research has been supported by an Australian Postgraduate Award Scholarship, a UWA Top-Up Scholarship and Travel Grants from the Graduate Research School and the School of Psychology at UWA.

I have been lucky to share space with some brilliant PhD students who have been a daily source of entertainment and support. Thank you to Suzanna, Emma, Adelln, Steph and Sylvie.

Thank you to my girlfriends, Saruchi, Danielle, Alysia, Kellie, and Nat. I am so very lucky to have you all in my life.

Finally, I want to express my deepest gratitude to my partner Colin, and my family; Mum, Dad, Danielle, and Jaymon. Thank you for celebrating every milestone with me, but more importantly, thank you for encouraging me during the tough times. I love you all, and I could not have done this without your love and support.
Section One

General Introduction
Chapter One

General Introduction

Clinical significance methodology (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991) provides a way of assessing the meaningfulness of a treatment outcome by considering the reliability of a change in the context of a patient’s resemblance to the normal, functional population. It has been used in clinical trials to measure the effectiveness of treatments (Asarnow et al., 2005; Bodenmann et al., 2008; Fisher & Wells, 2005; Forman, Herbert, Moitra, Yeomans, & Geller, 2007; Teri, Logsdon, Uomoto, & McCurry, 1997; Webster-Stratton, Hollinsworth, & Kolpacoff, 1989) and to guide clinical decisions such as the escalation or cessation of treatment (Lambert, Gregersen, & Burlingame, 2004). However, there are several assumptions that have been made regarding the use of clinical significance methodology, which remain largely untested.

- Firstly, it is a requirement of most clinical psychology journals that the clinical significance of a change is reported, yet the method of determining clinical significance is not always the same; nor is the outcome measure from which to measure change. This publication requirement thus makes the implicit assumption; perhaps incorrectly, that the calculation method and outcome measure used to evaluate the clinical significance of a change do not impact the result.

- Secondly, clinical significance categorisations are assumed to be trustworthy, despite limited evidence for the validity of interpretations made based on these evaluations. For instance, it is not known if a change evaluated as ‘clinically significant’ by one of the several classification methods translates to change that is meaningful either to the patient or the treatment provider.
• Thirdly, some have suggested that clinical significance methodology does not accommodate changes experienced by patients who begin treatment in the severe ranges (Jacobson & Revenstorf, 1988; Lambert, Hansen, & Bauer, 2008; Tingey, Lambert, Burlingame, & Hansen, 1996). In other words, clinical significance categorisations are used with the implicit assumption that the categories are equally appropriate to people at every level of severity and hence it has yet to be shown that the clinical significance methodology is sufficiently flexible to account for meaningful change experienced by different patient groups.

• Finally, through the language used, the implicit assumption has been made in clinical psychology that evaluations of clinical significance are stable, not dynamic, and as a result recovery is conceptualised as an outcome to be reached, as opposed to a process to be experienced.

The present dissertation will highlight the importance of testing each of these often implicit assumptions, it will examine each assumption, and it will provide recommendations as to how to accurately employ and interpret clinical significance methodology to achieve meaningful clinical goals. Testing these assumptions is imperative given the importance of the clinical judgments and decisions made based on evaluations of clinical significance.

Before each of these assumptions can be addressed in more detail, it is necessary to discuss the importance of evaluating outcomes in mental health, and the history of assessing outcomes using the clinical significance concept.
The Importance of Evaluating Outcomes

Outcome evaluation in mental health refers to the process of assessing and interpreting a patient’s functioning following mental health treatment, with the overarching goal of improving the effectiveness of treatment (Kazdin, 1994). Treatment efficacy refers to the degree to which a treatment is successful in carefully controlled clinical trials, while effectiveness refers to the degree to which a treatment is successful in real-world clinical practice (Brook & Lohr, 1985). Both forms of evaluation are an important part of science-informed practice (Page & Stritzke, 2006) and allow treatment providers to compare outcomes for different interventions, different patient populations, and different treatment agencies. These evaluations inform clinical decision making, with the ultimate aim to improve treatment outcomes for patients.

Methods of Outcome Evaluation

Evaluation of change, by definition, requires observations from a minimum of two points in time. In the context of mental health, the most practical measurement time-points are at pre-treatment and post-treatment. This allows for inferences to be made regarding changes in symptoms that occur at the same time as treatment is provided. Of course, factors other than the treatment itself can have an influence on changes occurring during this time; therefore causal interpretations must be made cautiously.

Outcomes can be evaluated by noting the presence or absence of a mental health diagnosis through a clinical interview (Elkin, 1994), however this is time-consuming, and provides only a categorical ‘yes/no’ response. The difference between two scores on a questionnaire measure, ideally recorded at pre-treatment and post-treatment, provides a richer outcome by virtue of using a continuous measure of change. However, a measure of the statistical significance of the difference between pre and post-treatment
scores is required in order to assess the likelihood that a change on a continuous measure has occurred by chance (Kazdin, 1994). Applying a test of significance increases the confidence that an observed difference is larger than that expected from measurement error, but statistical significance is not without its weaknesses. Selection of an inappropriate alpha value can introduce bias, and weak power can be an issue if the sample size is too small. Most importantly, a treatment effect may be “statistically significant” in cases where actual change is small, but not clinically meaningful (Nelson & Allred, 2005), especially when studies use large sample sizes.

One response to the problems with statistical significance has been to use effect sizes to evaluate outcomes. These permit judgements about the strength of the mean effect of a treatment and comparisons of the strengths of different treatment effects. An effect size, which is equivalent to a Z-score, is calculated by taking the difference between the mean of the experimental group and the mean of the control group and dividing this by the standard deviation. Cohen (1988) has suggested guidelines for interpreting effect size strength (e.g., 0.8 = large effect) and although some critics believe they lack empirical justification, no alternative guidelines have been presented (Olejnik & Algina, 2000).

Nonetheless, even though effect sizes address some of the issues with a sole reliance on statistical significance, they do not consider the extent to which a patient still requires treatment. That is, a treatment that demonstrates a large effect size may not necessarily correspond to meaningful changes in a person’s daily life. For instance, it could be conceivable for a weight loss program with obese individuals to find a large effect size based on a decrease of 5kg to 142kg for its experimental group, compared to 0kg weight loss in the control group. However if the mean height was 177cm, then the individuals would still be in the “super obese” category with the attendant health risks.
Furthermore, both statistical significance and effect size measures assess the average outcome for a group of patients. This is potentially problematic since although a patient group at post-treatment may show a statistically significant change in symptoms, and the treatment may yield a large effect size, there may still be individuals who have worsened as a result of the treatment. Examination of the mean change alone does not reflect how many people showed the desired effect. That is, the same mean change could arise because everyone shows the same level of change, or because there are a couple of patients with dramatic improvement while most patients show little or no change. Thus, in certain contexts, measures of individual change are preferred because they will identify those individuals who have worsened, or remained unchanged as a result of treatment, and therefore be able provide them with an opportunity for further treatment. Individual patient change can be calculated using an individual effect size (pre-treatment score minus post-treatment score, divided by the standard deviation for matched patients) allowing for a comparison between the individual patient effect size and effect sizes published in the literature. Although this method could capture whether an individual change was better than expected compared to findings in the literature, it does not provide information about the meaningfulness of the change. *Clinical significance* methodology overcomes both of these preceding issues by allowing treatment providers to assess and describe the practical meaning of a treatment outcome for an individual patient. Group treatment outcomes can still be reported, but instead of evaluating treatment with reference to a statistical significance value or an effect size, the proportion of patients achieving clinically significant change is reported.
What is Clinically Significant Change?

A change is considered *clinically significant* if it is meaningful, or important in a clinical sense. In earlier times, clinically significant changes were described as having ‘social validity’ (Kazdin, 1977; Wolf, 1978); that is, being considered important from a societal perspective. For example, a socially valid outcome as a result of an alcohol abuse treatment group would be for the patient to cease drinking. Other outcomes having a socially valid impact would include reduced mortality, hospital admissions, days of absence from work, or crime rates. Some have argued that if a social impact of a treatment is not observed, then the treatment cannot be considered truly effective, however it emerged over time that socially valid outcomes of treatment were not always easily identifiable. In addition, some social variables can be insensitive as measures of treatment outcome, and it is impossible to determine whether the treatment received has had a sole, direct impact on the relevant social variable (Kazdin, 1994). For a review of the history of social validity issues see Ogles, Lunnen, and Bonesteel, 2001.

Clinical significance has also been evaluated subjectively. Evaluations of clinical significance can be made based on the existence of a certain disorder (e.g., diagnostic criteria for mental disorders in the DSM-5 (American Psychiatric Association, 2013) usually include “clinically significant impairment”) or the change made during various stages of treatment. For example, the practical importance or meaningfulness of a change made during treatment can be judged by the treatment provider, the patient’s significant other, or the patient themselves. Early in the 20th century, subjective rating scales were used by treatment providers in an attempt to assess the clinical significance of outcomes. Options for categorising each patient’s change included ‘cured’, ‘much improved’, ‘slightly improved’, ‘improved’, and ‘unimproved’ (Bergin, 1971). These ratings were based on the treatment provider’s
opinion of the importance of the change, since the aim from a treatment provider’s perspective is to ‘cure’ the disorder that the patient initially presented to treatment with. It may be argued that the practical importance of a change as a result of treatment cannot be evaluated without including judgments from the patient and their loved ones, since each patient’s measure of the ‘importance’ of their change will be based on a set of factors unique to them and their environment. The inevitable downside of having such flexibility in evaluations of the importance of the change is that they are susceptible to bias and therefore may not be psychometrically sound (Kazdin, 1994).

A clinically significant change has also been defined as a return to normal functioning, or a departure from ‘non-normal’ functioning. In many areas of measurement there are widely held, objective standards regarding what is normal. For example, fitness trainers can assess a person’s health based on whether their weight falls into the ‘healthy’ (normal) range, psychologists can judge a child’s intelligence on whether their IQ falls in the ‘average’ range, and doctors can judge the health of a patient’s liver based on whether results from liver function blood tests fall into the ‘normal’ range. These agreed-upon standards are based on data from ‘normally-functioning’ members of the population. In mental health, when normative data exists for a particular phenomenon (e.g., the degree to which depressive symptoms are present in the normal population, and in a patient population) then the clinical significance of a change made by a patient treated for depression can be evaluated in two ways. If a patient has moved markedly out of the ‘dysfunctional’, patient range, or moved markedly into the ‘functional’, normal range then this is considered a clinically significant change. Problems arise, however, when normative data are not available. In addition, how is ‘markedly’ defined?
Kendall and Grove (1988) described a view of normal functioning in which past patients are not distinguishable from their non-treatment seeking peers. Alternatively, diagnostic interviews can be used to assess a patient’s return to normal functioning. By conducting interviews at pre-treatment, when it would be expected that a patient would meet diagnostic criteria for a disorder, and again at post-treatment, the clinical significance of a patient’s change can be assessed (Ogles, Lunnen, & Bonesteel, 2001).

**Jacobson-Truax Method for Evaluating Clinical Significance**

Jacobson, Follette, and Revenstorf (1984; and later Jacobson & Truax, 1991) understood the criticism directed towards the use of statistical significance and effect sizes to evaluate treatment outcomes, and proposed the concept of clinical significance to provide a useful way to evaluate the meaningfulness of an individual’s change. Furthermore, they acknowledged the need to clearly define clinical significance, and using agreed-upon conventions, categorise change into a range of possible outcomes, not simply a dichotomy (i.e., clinically significant change versus not clinically significant change) which would lack specificity. Jacobson et al. (1984) argued that two criteria needed to be considered in an evaluation of the clinical significance of a change: (a) whether the patient resembles a member of the functional population or the dysfunction population, and (b) whether the change made during treatment has been reliable. To evaluate the clinical significance of a change based on these criteria, they needed to be quantitatively defined.

Firstly, three types of cut-off points separating the dysfunctional from the functional population were proposed. The functional population is conceptualised by Jacobson and Traux (1991) as those individuals who are not involved in therapy, whereas the dysfunctional population represents those individuals who are. When symptom severity of these two distributions is plotted, there will necessarily be some
overlap represented by two groups: (a) individuals who have low symptom severity and seek treatment; and (b) individuals who have high symptom severity and do not seek treatment. Therefore they are not distinct populations. Cut-off A is a point two standard deviations away from the mean of the dysfunctional population, and passing this cut-off represents a movement out of the dysfunctional range. It is recommended that this cut-off be used when data from a normative sample are not available. Cut-off B is a point two standard deviations away from the mean of the functional population, and passing this cut-off represents a movement into the functional range. Jacobson and Truax (1991) suggest that cut-off B should be used when the functional population and dysfunctional population do not overlap, since this will then only capture those patients who are simultaneously in the functional range and out of the dysfunctional range. Cut-off C is a mid-point between the two distributions, and represents the point at which the proportion of true members of the functional population who are misclassified as dysfunctional is equal to the proportion of true members of the dysfunctional population who are classified as functional, when the entire population is classified (Hsu, 1999). It is recommended that cut-off C be used when normative values are available. Means and standard deviations for the functional population need to come from normative samples (e.g., multiple community samples, often found in the outcome measure manual) measured using the relevant outcome measure in the relevant population. These will therefore be consistent when the same outcome measure is used for the same population. Means and standard deviations for the dysfunctional population come from a sample that resembles the patient sample being evaluated. This may be the current pre-treatment patient sample, or another sample of similar demographic, diagnosis, and symptom severity. It is important to note that the functional and dysfunctional distributions cannot be universally and clearly defined as they will vary substantially between cultures, health systems, diagnoses, educational levels and so on. Therefore
they are considered arbitrary categories, unique to the population of interest. Of course, it is important to reflect on the potential influence that the choice of parameters can have on the nature of the distributions created. For example, if the functional distribution used in one study consists of an “asymptomatic” sample, containing only individuals who report no symptoms, and another consists of a “community” sample, naturally containing some individuals who report symptoms, then it will be more difficult to move into the functional distribution for those people assessed using the “functional” distribution represented by “asymptomatic” individuals. Ultimately, the choice of sample used to represent the functional distribution should rest upon what is deemed appropriate given the current monitoring application or research question of interest.

Secondly, reliable change needs to be defined. This criterion is necessary since it is possible for a patient to move from the dysfunctional range to the functional range when they have changed by an amount on a measurement scale that may reflect measurement error only. Jacobson et al. (1984) presented a Reliable Change Index (RCI) which takes into account the change made by the patient on the relevant outcome measure, in the context of the standard error of measurement. The RCI is the single-subject design counterpart to a conventional group-based test of statistical significance, that is, a paired $t$-test, which assesses whether the change in group mean is considered statistically significant (Jacobson & Revenstorf, 1988). It is calculated with the following formula:

$$RCI = \frac{(x_2 - x_1)}{\sqrt{2 \times SE^2}}$$

where:

$$SE = SD_1 \times \sqrt{(1 - R)}$$
and:

\[ x_1 = \text{patient pre score} \]
\[ x_2 = \text{patient post score} \]
\[ SD_1 = \text{standard deviation of all pre scores} \]
\[ R = \text{reliability of measurement instrument} \]

The term ‘reliable change’ is a function of the reliability of the outcome measure and the standard deviation of the population. When an RCI is greater than 1.96 or less than -1.96 it reflects a reliable change, corresponding to an alpha level of 0.05. If a measurement instrument has low reliability, a large amount of change is required before the change is considered reliable, resulting in lower improvement rates as demonstrated by Speer (1992). Conversely, for measurement instruments with high reliability, a smaller amount of change will be considered reliable, and greater improvement rates will result (Speer, 1992). Therefore, the reliability of the outcome measure used to assess clinically significant change can substantially impact on clinical significance classifications. The same can be said for the reliability coefficient that is used in calculating the RCI, since the lower the reliability score of a measure, the more difficult it will be for a patient’s change to be considered reliably changed. Initially in the clinical significance literature, the test-retest reliability score was recommended as this was considered best at assessing the change that occurs in patients not receiving treatment (e.g., Jacobson, Follette, & Revenstorf, 1984; Jacobson & Revenstorf, 1988; Jacobson & Truax, 1991). However, after further consideration, of the practical difficulties in using test-retest reliability, it was recommended that the internal consistency (alpha) reliability coefficient be used instead (Martinovich, Saunders, & Howard, 1996; Tingey, Lambert, Burlingame, & Hansen, 1996). Martinovich et al. stated that since some patients will respond to treatment and others will not, test-retest reliabilities will be deflated by real individual differences in treatment response during
the first two weeks. This therefore results in smaller observed changes being interpreted as reliable, although individual differences in change are less reliable. In addition, as discussed by Tingey et al. (1996) certain populations such as inpatients can display erratic behaviour patterns that could affect reliability estimates, especially in the two week period prior to a treatment commencing, in which test-retest reliability is estimated. Based on these issues, the reliability coefficient of internal consistency; Cronbach’s alpha, is used in the calculation of reliable change in this thesis. The source of the internal consistency figure should be reported, whether this is measured in the current study sample or taken from the existing literature (e.g., measure development studies or measure evaluation studies).

These two systems for categorising patient change result in four possible outcomes (Jacobson & Truax, 1991): recovered, in which a patient has moved reliably from the dysfunctional population to the functional population; improved, in which a patient has made a reliable change in a positive direction but does not yet resemble a member of the functional population; unchanged, in which no reliable change has been made by the patient; and deteriorated, in which a patient has made a reliable change in a negative direction. The proportion of patients assigned to each category based on their responses on a relevant outcome measure following mental health treatment can then be used to evaluate the extent to which a treatment outcome was successful. For example, a treatment could be evaluated as being more successful than another if it had higher recovery rates or lower deterioration rates; depending on the way in which success is operationalized.

It could be argued that by placing patients into clinical significance categories based on their symptom severity scores, rich information is being lost, since a continuous score is ‘reduced’ to a category. However, this view neglects to consider the
important function of clinical significance classifications: they serve to classify a patient’s change during treatment, not solely their post-treatment score. For example, on for a symptom measure where patients can score between 0 and 20, there are 21 possible post-treatments scores they could have. However, there are 441 (21 x 21) possible combinations for pre-treatment to post-treatment change. It is conceptually simpler for clinicians to group these 441 different changes into a fewer number of meaningful and easily understood categories. The methodology allows clinicians to easily recognise what ‘type’ of change their patients have made, and using this information to measure treatment effectiveness hence guiding their intervention.

An Assumption that the Approach Taken to Assess Clinical Significance Does Not Matter

Due to the mounting evidence pointing to the usefulness of evaluations of clinical significance in making clinical decisions and interpreting the degree to which a patient has responded to treatment, it would make sense for journals to make explicit requests for authors to report the clinical significance of results submitted for publication. However in the most recent CONSORT statement (Moher et al., 2010) which provides guidelines for conducting randomised controlled trials, clinically significant change is alluded to only twice. On page 8 it says: “a study should be large enough to have a high probability (power) of detecting as statistically significant a clinically important difference of a given size if such a difference exists” and on page 20 it reads: “the difference between statistical and clinical importance should always be borne in mind”. While these points are well-made, the absence of any clear specification of clinical significance means that their interpretation is open to subjectivity and bias. This lack of importance placed on evaluations of clinical significance in randomised
controlled trials is concerning, given the disadvantages of relying solely on measures of statistical significance and effect size to evaluate treatment outcomes.

Furthermore, following a search of journals belonging to the field of psychology from Wiley Online Library, Elsevier, SAGE Online, and Taylor & Francis Online, as well as all APA journals, only two journals were identified that made recommendations to authors regarding reporting clinical significance. The *Journal of Consulting and Clinical Psychology* states that it “requires the statistical reporting of measures that convey clinical significance”, in a similar manner to the *Journal of Clinical Child and Adolescent Psychology* that states “authors should include indicators of clinically significant change”. This existence of these guidelines at all is encouraging, but these guidelines then specify that “authors may use one of several approaches that have been recommended for capturing clinical significance”. The implicit assumption that underlies this statement is that the approach used to capture clinically significant change does not matter. The guideline then suggests that clinical significance approaches include “the reliable change index”, “the extent to which dysfunctional individuals show movement into the functional distribution”, or “other normative comparisons”. Confusingly, these approaches to measuring clinically significant change that follow this recommendation, *on their own*, do not represent clinically significant change as it has been defined in the literature (Jacobson et al., 1984; Jacobson & Truax, 1991). A change can only be clinically significance if the change is reliable, *and* the individual has moved from the dysfunctional range into the functional range. As Jacobson and Revenstorf (1988) state: “it is only when reliable change is used in conjunction with the cut-off point between the functional and dysfunctional distributions that it has any relevance to the determination of clinically significant change” (p136). The lack of reporting guidelines provided by the vast majority of psychology journals, as well as the unclear descriptions given in the few journals that do provide such guidelines, is
making, perhaps incorrectly, the assumption that the approach used to capture clinically significant change does not affect the evaluations made.

Making such an assumption is unwise since there are several factors that can vary when evaluating the clinical significance of an outcome. Firstly, the choice of clinical significance calculation method can vary (comprised of both a reliable change index, and a cut-off point separating the functional and dysfunctional distributions), and secondly, the choice of outcome measure used to measure change in patients can vary. Past research implies that the clinical significance of an outcome could vary based purely on the way in which clinical significance is evaluated. Fisher and Wells (2005) compared the effectiveness of various treatments for Obsessive Compulsive Disorder and found that rates of recovery for patients undergoing Exposure and Response Prevention (ERP) varied between 24% (Van Oppen et al., 1995) and 73% (Franklin, Abramowitz, Kozak, Levitt, & Foa, 2000). At first glance, these divergent results could imply wide variation in outcomes for the treatment. However, there are several possible reasons why this difference may be so large. Firstly, there is a 4-point difference between the cut-off chosen by the two studies to separate the dysfunctional population from the functional population on the Y-BOCS (Yale Brown Obsessive Compulsive Scale; Goodman et al., 1989). Franklin et al. used a cut-off of 16 and van Oppen et al. used 12, which meant that the probability of entering the healthy range was greater in the former study solely because of the decision to use a less stringent value.

Different Clinical Significance Calculation Methods. The Jacobson-Truax (1991) approach to defining reliable change considers the actual difference between pre- and post-treatment scores in the context of the standard error of measurement. Critics of this method suggested that it does not take into account regression to the mean; a phenomenon which describes the tendency for more extreme scores (e.g., more severe
symptoms) to become less extreme over time (Campbell & Kenny, 1999). Speer (1992) suggested that the greater degree to which patient scores deviated from the functional population at pre-treatment, the greater the chance of regression to the mean occurring. As a result of this criticism, several new methods were proposed including the Gulliksen-Lord-Novick method (Hsu, 1989, 1995), the Nunnally-Kotsch method (Nunnally & Kotsch, 1983), the Edwards-Nunnally method (Speer, 1992), and the Hsu-Linn-Lord method (Hsu, 1989). Proponents of each method proposed that regression to the mean was accounted for within their unique method for calculating reliable change. Hageman and Arrindell (1999) also aimed to correct for regression to the mean by presenting an alternative method for calculating reliable change, which incorporated weighted reliability values for the measure at pre- and post-treatment. However, unlike the other methods, which recommended the use of Jacobson and Truax’s (1991) cut-off points, Hageman and Arrindell modified the formula for determining the cut-off point separating the functional from the dysfunctional population. They claim that rather than calculating cut-offs based on observed scores from the functional and dysfunctional population, it is more accurate to calculate cut-offs based on ‘true score equivalents’. The cut-off separating these two populations from one another has an important influence over whether a change is considered clinically significant: it captures the “meaningfulness” aspect of the conceptualisation of clinically significant change. Specifically, this criterion separates those patients who have made a reliable, positive change into the two categories: improved (i.e., those who remain in the dysfunctional population at post-treatment), and recovered (i.e., those who have entered the functional population at post-treatment). Several research studies compare clinical significance evaluations based on different calculation methods (Atkins, Bedics, McGlinchey, & Beauchaine, 2005; Bauer, Lambert, & Nielsen, 2004; McGlinchey, Atkins, & Jacobson, 2002; Speer & Greenbaum, 1995), which implies that this issue is believed to be of
importance in the outcomes monitoring literature. Yet, the choice of which method to use remains uncertain. In some studies, the Jacobson-Truax method is used only because it is the “most commonly used method”; for example in Brouwer, Meijer, and Zevalkink (2013). This strengthens the current assumption that the calculation method used does not matter. Another example of this occurs in Gonda, Deane, and Murugesan (2012) in which the rationale for using the Jacobson-Truax method over other methods is that it “seems particularly useful” (p651) with the justification given applying equally to four other published methods for calculating clinical significance. The issue of calculation method will be addressed in more detail in Chapter 2.

**Different Outcome Measures.** Scores on outcome measures at pre- and post-treatment form the difference score from which the assessment of reliable change is made. The choice of outcome measure to use to assess clinically significant change should depend on the primary treatment goal. Some outcome measures are patient-rated, for example, self-report mood scales such as the Beck Depression Inventory (BDI; Beck, Steer, & Carbin, 1988), and some are clinician-rated, for example, the Health of the Nation Outcome Scale (HoNOS; Wing et al., 1998). Of course, the suitability of an outcome measure for evaluation purposes is contingent on the domain of interest (e.g., quality of life, social functioning, depressive symptomatology or satisfaction with care). In contrast to studies comparing the effects of using different clinical significance evaluation methods, there exist fewer studies exploring the effects of using different outcome measures on evaluations made regarding clinically significant change (Beckstead et al., 2003). This may either imply that choice of outcome measure is not seen as an important issue as it is predicted to result in similar evaluations, or that it will make a difference but that users of clinical significance methodology will know this, and therefore choose their outcome measure wisely. Moreover, it is not known what effect using different combinations of outcome measures and clinical significance
classification methods will have on evaluations made regarding the extent of clinically significant change observed in a patient group. After all, both choices need to be made, and employed simultaneously in order to produce a clinical significance classification. This topic will be explored in Chapter 2 of this dissertation. The greater extent to which clinical significance classifications are consistent when different reliable change formulas, cut-off points, and outcome measures are used, the more comparable, and therefore clinically useful, these classifications will be. It would then follow that journals making recommendations to report clinical significance classifications indeed do not need to specify how such evaluations are made. But, where inconsistencies exist, recommendations need to be made regarding the most accurate way to classify outcomes according to clinical significance. The issue of the impact of measurement instrument on clinical significance outcomes will be addressed in more detail in Chapter 2.

An Assumption that Clinical Significance Classifications are Equally Valid

For argument’s sake, let us assume that the approach taken to assess clinically significant change does not matter, and that classifications are consistent, regardless of the approach. The resulting labels of recovered, improved, unchanged and deteriorated assigned to patients following treatment will be meaningless if the classifications are not ecologically valid or meaningful in a real-world sense. If a patient classified as recovered does not display characteristics expected from a patient who has recovered from a mental illness (e.g., improved quality of life, improved job functioning, lower self-reported symptom severity), then what is the advantage of assigning them with this label at all? Similarly, if a patient is classified as deteriorated according to an assessment of clinically significant change following treatment, but they do not show signs related to what would be expected from a patient who has worsened, then a label
is of no practical or clinical use. More importantly, clinical decisions made based on clinical significance approaches that do not show ecological validity might result in adverse outcomes. For example, a patient classified as recovered by an invalid approach could have their treatment terminated before they have benefitted from it. Conversely, a patient who is classified incorrectly as unchanged may have actually benefitted sufficiently from treatment, but would unnecessarily expend resources if treatment were continued.

The approach taken to address the issue of ecological validity depends on a broader issue: what is “recovery”? By some in the mental health field it is considered to be an outcome that can be reached as the result of treatment. When defined in this way, recovery usually takes into account the extent to which a patient experiences the symptoms of their mental illness at the end of treatment (Schrank & Slade, 2007). However, “recovery” is not always defined in this way. By others, it is considered to be a process or journey that is constantly being experienced, both during treatment and following treatment. When recovery is conceptualised as a journey, the definition usually includes elements such as finding meaning in life, hope, control, opportunity, and goals (Davidson, O’Connell, Tondora, Lawless, & Evans, 2005). Hence, it is not clear the extent to which the classifications of “recovered” obtained when using clinical significance map onto the meanings that different groups give to the concept of recovery.

It is clear that the ecological validity of evaluating outcomes using clinical significance methodology is an important issue. Several findings related to validity have been published, for example Ankuta and Abeles (1993) found that patients who made a clinically significant change on the Symptom Check-List 90-Revised (SCL-90R; Derogatis, 1977) reported greater levels of satisfaction with treatment than those who
did not make a clinically significant change. Since it cannot be concluded that satisfaction results only from a true clinically significant change (i.e., it is possible for a patient to feel unsatisfied with treatment despite achieving a positive outcome and vice versa), this finding provides weak support for the validity of clinical significance evaluations. Additionally, convergent validity of clinical significance classifications with related scales has been assessed by Ogles, Lambert, and Sawyer (1995), Lambert, Okiishi, Finch, and Johnson (1998) and Newnham, Harwood, and Page (2007). Results largely suggest that clinically significant change as classified using the Jacobson-Truax method is related to improvements on other related measures; however it is not known whether these findings extend to other clinical significance methods. The continued use and recommendation of clinical significance methodology assumes that clinical significance classifications are valid; yet better evidence supporting their validity is required. What is needed is a systematic exploration of the associations between clinical significance classifications and ‘real-world’ variables that are considered vitally important in relation to recovery versus no clinically significant change, following treatment. Such an exploration will require a combination of subjective and objective measures of importance to patients and treatment providers. This exploration is carried out in Chapter 3 of this dissertation.

As mentioned earlier in this chapter, within the guidelines for reporting clinically significant change in articles submitted to Journal of Consulting and Clinical Psychology incorrectly mentions the two components of evaluating clinical significance separately as “indicators of clinically significant change”: the reliable change index, and the extent to which a patient has moved from the dysfunctional population to the functional population. Several studies have used one of these alone as a measure of clinically significant change (e.g., Lunnen & Ogles, 1998; Speer & Greenbaum, 1995; Wise, 2010). Although this is not correct according to Jacobson and Truax’s (1991)
definition of clinically significant change, which requires both a reliable change and a movement towards normality, these studies raise an important question: which component of evaluating clinically significant change holds the most ecological validity? That is, does recovery in a ‘real-world’ sense relate more to the magnitude and reliability of a change made on a relevant outcome measure, or the distribution that a patient belongs to at the end of treatment? This issue is explored in Chapter 4.

An Assumption that Clinical Significance Evaluations are Equally Appropriate for Individuals at Every Severity Level

For patients who experience particularly severe symptoms, achieving *recovery* may not be a fair and realistic goal. Since clinical significance methodology requires a reliable move from the dysfunctional range to the functional range for a change to be considered clinically significant, patients falling in the tail end of the dysfunctional distribution will need to improve by a much larger magnitude than patients experiencing mild to moderate symptoms. Clinical significance methodology has been criticised for not recognising meaningful changes that might be made in the lives of severely symptomatic patients during treatment (Jacobson & Revenstorf, 1988). For example, a patient experiencing schizophrenia might improve as a result of treatment to the point at which they are able to hold down a casual job, yet they will not resemble a member of the functional population in terms of the severity of the symptoms they still experience. Others have argued that using only two extreme distributions to characterise severity (functional and dysfunctional) incorrectly assumes that the population forms a bimodal distribution (Wampold & Jenson, 1986). Despite these criticisms, no clear recommendations have been made that advise against their use for patients experiencing severe symptoms, which assumes that *clinical significance categories hold the same meaning for individuals at every severity level*. 
Tingey, Lambert, Burlingame and Hansen (1996) recognised these criticisms towards clinical significance methodology and agreed that categorisations of clinically significant change needed to be flexible enough to adapt according to the severity of the group being evaluated. They argued that the purpose of assessing clinically significant change was to identify those individuals who have experienced change that is important and meaningful, and that this could occur without movement into the ‘normally functioning’ population. Tingey et al. (1996) proposed an extension to traditional clinical significance methodology whereby multiple distributions were used to characterise individual change, based on symptom severity as measured by the SCL-90R (Derogatis, 1977). Distributions included “asymptomatic” (individuals identified who did not experience symptoms), “community” (members of the population not currently seeking mental health treatment but who may or may not be experiencing symptoms), “outpatient” (individuals attending an outpatient mental health service), and “inpatient” (individuals attending an inpatient mental health service). As expected, each of these four groups had a mathematically distinct distribution of symptom severity, in the same way that traditional models of clinical significance characterise two distinct distributions: functional and dysfunctional. However, in contrast to traditional models whereby clinically significant change can only occur if an individual has moved into the functional distribution, Tingey et al. argued that change can be clinically significant if an individual has moved into an adjacent and less severe distribution. In this way, a patient who moves from the inpatient distribution to the outpatient distribution is considered to have made a clinically significant change, although they are still experiencing symptoms to a degree that places them in a ‘dysfunctional’ population.

This new conceptualisation of clinical significance takes into consideration an important goal of clinical significance methodology: to characterise change that is ‘meaningful’ (Jacobson & Truax, 1991). Change can be meaningful to both the patient
and their treatment provider. For a patient engaged in inpatient care (i.e., involving overnight stays in hospital), the goal from both the perspective of the patient and the treatment provider is for this patient to be well enough to leave hospital. It may not be that this patient resembles a member of the normal, functioning population at post-treatment, and therefore when using the traditional clinical significance classification methods the patient would not be categorised as recovered, but perhaps improved instead. They may still require outpatient care (e.g., attendance of appointments during the day) in the weeks and months following cessation of inpatient care. However, as argued by Tingey et al. (1996), a change such as this should still be considered clinically significant, as it is clearly meaningful to both the patient and the treatment provider. Therefore, making the assumption traditional clinical significance categories are equally appropriate for all patients could mean that important change is not recognised; a possibility that is particularly risky when clinical decisions are made based on whether or not clinically significant change has occurred.

Making this assumption is also risky when rates of clinically significant change are compared across treatment sites that may specialise in treating patients of different symptom severities. In a study by Hansen, Lambert and Forman (2002), rates of clinically significant change were compared across six different mental health treatment sites. They found that rates of recovery varied between 8.6% (state community mental health) and 20.0% (training community mental health). On a cursory examination it may appear that the most effective site was the training community mental health clinic, however, due to the nature of training clinics, it is likely that the average patient seen at this site is of a higher functioning than the average patient seen at the state community mental health site (and hence more likely to enter the “healthy” range after successful treatment). Therefore, conclusions about the relative effectiveness of each site become very difficult to interpret.
Although the work of Tingey et al. (1996) is well placed to overcome this issue, there has been little effort to replicate or extend their work by applying it to other outcome measures or patient groups. Due to increasing recommendations of the use of clinical significance methodology that do not discuss its appropriateness for extreme groups, it is important that research demonstrates not only that clinically significant change is less achievable for patients with severe symptoms, but also that clinical significance methodology can be flexible in accounting for meaningful change in different patient groups. This will be the focus of Chapter 4 of this dissertation.

An Assumption that Clinical Significance Methodology Considers Classifications as Outcomes Rather than Processes

In the mental health literature, recovery has been conceptualised both as an outcome and as a process (Schrank & Slade, 2007). Those who view recovery as an outcome usually define recovery as the absence of symptoms, whereas those who view recovery as a process usually define recovery as the presence of factors such as hope, opportunity for growth, control over one’s life, and goals for the future (Davidson, Drake, Schmutte, Dinzeo, & Andres-Hyman, 2009). Considering recovery as a process implies that it can continue despite the discontinuation of treatment. In effectiveness studies in clinical psychology it is usual to measure the effectiveness of treatment at post-treatment but also at a follow-up time, to determine whether changes made during treatment are maintained. Research suggests that scores on symptom measures at follow-up are generally stable (Lambert & Bergin, 1994). Since clinical decisions are made based on clinical significance classifications as well as conclusions drawn regarding treatment effectiveness, this assumes that clinical significance classifications are considered to be outcomes to be reached, as opposed to processes to be experienced.
However, when clinical significance classifications are considered to be outcomes rather than processes, treatment providers’ conceptualisation of recovery is narrowed. It is possible for two treatment outcomes to appear similar when assessed at post-treatment but to experience different forms of change subsequent to this, and perhaps depending on the form of treatment received. A study by Mattick and Peters (1988) compared the effectiveness of using guided exposure to treat social phobia, both with and without cognitive restructuring, at post-treatment and three months later. The combined treatment was slightly more effective than the other treatment at post-treatment; however the effectiveness of the combined treatment continued to increase in the three months following treatment, whilst the effectiveness of guided exposure alone remained stable. Although the added benefit of cognitive therapy for social phobia has not always been replicated, this finding illustrates that improvement can continue following the end of treatment, presumably when patients are equipped with particular skills during treatment that they can apply to their everyday life.

If clinical significance classifications such as recovery, improvement, no change, and deterioration are conceptualised as processes rather than outcomes, this may allow assessments of categorisations at post-treatment to predict future change, or lack of change following the cessation of care. Viewing classifications as processes may also decrease the sense of hopelessness associated with less desirable outcomes such as no change or deterioration. For example, if it is found that some patients classified as unchanged at post-treatment can actually achieve a reliable improvement following the end of treatment, this provides hope to treatment providers as well as patients, and perhaps certain measures can be taken to increase this patient’s chances of improving post-treatment. Chapter 5 of the present dissertation will examine the stability of clinical significance classifications in the six weeks following post-treatment and assess the assumption that changes occurring during treatment are best conceptualised as outcomes.
or processes. In addition, Chapter 5 will explore whether common patterns of change (either improvement or deterioration) or stability are evident in patients assessed at post-treatment and then six weeks later, with the aim of making recommendations regarding the predictive value of clinical significance classifications at post-treatment. This may allow the opportunity to improve the success of treatment even after it has reached its conclusion.

**Thesis Aims and Outline**

As demonstrated, clinical significance methodology is an important aspect of outcome monitoring in mental health. This method of evaluation is commonly used to guide clinical decisions as well as evaluate the effectiveness of treatments, with the ultimate aim of improving treatment outcomes for patients. Therefore it is alarming that several assumptions regarding clinically significant change have remained untested, despite the increases in its use. This dissertation is comprised of two published papers, and three papers in preparation for journal submission. Some chapter forewords are included which contain important findings that link chapters together to form a conceptual whole.

- *An assumption that the approach taken to assess clinical significance does not matter* is explored in Chapter 2 (paper published in *Clinical Psychology: Science and Practice*).
- *An exploration of the differences between patient perception of change and clinical significance classifications of change* is presented in the foreword to Chapter 3.
- *An assumption that clinical significance classifications yielded by different calculation methods are equally valid* is explored in the main body of Chapter 3 (invited submission for special issue of *Australian Psychologist* in preparation).
• An exploration of the contributions of different components of clinical significance classifications (i.e., reliable change index and the range patients belong to) is presented in Chapter 4 (paper in preparation).

• An exploration of the effect of patients’ initial symptom severity on likelihood of making clinically significant change is presented in the foreword to Chapter 5.

• An assumption that clinical significance evaluations are equally appropriate for individuals at every severity level is explored in the main body of Chapter 5 (paper published in Psychological Assessment).

• An assumption that clinical significance methodology considers classifications as outcomes rather than processes is explored in Chapter 6 (paper in preparation).

The final section of this thesis consists of a general discussion of the examination of each assumption, clear recommendations for future use of clinical significance methodology, and future research directions.

It is vital that these assumptions are carefully examined so that clearer recommendations can be made regarding their use. This will improve the value of clinical significance classifications to those who monitor outcomes in mental health, which ultimately has the potential to improve outcomes for patients.
References


Hsu, L. M. (1999). Caveats concerning comparisons of change rates obtained with five methods of identifying significant client changes: Comment on Speer and


Section Two
Reliability and Validity of Clinical Significance Classifications
Chapter Two

Chapter Two consists of a published paper:

How Consistent are Clinical Significance Classifications when Calculation Methods and Outcome Measures Differ?

Abstract

Outcome studies report the percentage of clinically significant outcomes; however, the reliability of these classifications is unclear. The current study explored the extent to which inconsistencies arise in classifying patient outcomes using five clinical significance calculation methods and three outcome measures. Adult inpatients (N=2676) treated for depression completed the three outcome measures pre- and post-treatment. Their outcomes were classified as recovered, improved, unchanged, or deteriorated using selected clinical significance calculation methods and outcome measures. The choice of outcome measure used in calculating clinical significance had a greater impact on the classification than the choice of clinical significance calculation method. Guidelines for reporting recovery rates are presented to ensure that appropriate conclusions are drawn.
Accurate interpretation of data regarding the success of a course of mental health treatment is vital in comparing the effectiveness of different treatments within treatment facilities, as well as the relative performances of individual staff or entire treatment facilities. Treatment outcome information, such as improvement rates, may signal when a patient is ready to cease treatment (Lambert, Okiishi, Finch, & Johnson, 1998). The importance of accuracy when making such choices is clear; when a patient undergoes treatment for longer than required, clinician time is expended unnecessarily. Instead, this resource could be more usefully allocated to a patient who is more ‘in need’. Conversely, it can be detrimental to a patient if treatment is ceased prematurely.

It is generally agreed that the optimal outcome in mental health treatment is that of recovery, a term that has been defined as an eradication of symptoms and amelioration of the deficits caused by an illness (Davidson, Drake, Schmutte, Dinzeo, & Andres-Hyman, 2009). For the purposes of monitoring patient treatment outcomes, recovery is most usefully described quantitatively, allowing for objective comparisons of treatment outcomes within and between patients. Currently, many quantitative methods exist in which to determine whether or not a patient has ‘recovered’ (Hageman & Arrindell, 1999; Hsu, 1989; Jacobson & Truax, 1991; Nunnally & Kotsch, 1983; Speer, 1992), each of which requires pre- and post-treatment scores from an appropriate outcome measure. Given the existence of multiple ways in which to classify a patient’s outcome, the key question arises: how consistent are such methods for calculating a patient’s outcome, and if there are inconsistencies, what effect could these inconsistencies have on decisions made regarding resource allocation, and relative effectiveness of treatments or treatment providers? To accurately answer this question, both (a) the method of calculating the clinical significance of a patient’s treatment outcome, and (b) the outcome measure that is administered pre- and post-treatment, need to be examined simultaneously.
Clinical Significance

A necessary condition for the evaluation of a mental health intervention is an optimal measure of pre- to post-treatment change. However, such a measurement is not sufficient for a comprehensive evaluation. A pre- to post-treatment effect size does not convey information about the nature of the outcome; specifically whether the person has moved into a healthy range. To address this weakness, outcomes have been categorised in terms of their “clinical significance”. Clinical significance methods typically classify patient outcomes on the basis of the strength of the effect size of the treatment outcome combined with an examination of whether the patient has reliably moved from the dysfunctional population to the functional population.

Several methods for calculating the clinical significance of a patient’s treatment outcome have been published. Arguably, the most popular of these methods is the Jacobson-Truax method (JT; Jacobson & Truax, 1991). The JT method classifies a patient’s outcome based on the reliability of the pre- to post-treatment change (i.e., does the change reflect more than the variations expected from an imprecise outcome measure?) and whether or not the patient has moved from the dysfunctional population to the functional population. By including thresholds to identify when a change is “reliable” and when the person is in the “functional” population distribution, it is possible to categorise outcomes into four clinically meaningful categories and respond to the criticism that effect sizes do not fully describe the nature of change. These categories are as follows: (a) recovered: a patient has reliably moved from the dysfunctional to the functional population; (b) improved: a patient has reliably moved towards the functional population but is not yet a part of it; (c) unchanged: a patient has
not made a reliable change; and (d) deteriorated: a patient has reliably moved away from the functional population.¹

Nonetheless, while clinical significance provides a method of categorisation of patient outcomes, it is not without its problems. Hsu (1989, 1995) argued that the JT method of calculating clinical significance did not consider regression to the mean. This phenomenon is the tendency for extreme scores on a particular measure to be followed by scores that are closer to the mean (Barnett, van der Pols, & Dobson, 2005). Regression to the mean will be more prevalent in samples in which patients score more extremely on a particular measure at pre-treatment. One example of this occurrence could be in efficacy studies. As efficacy studies deliberately select participants that meet certain criteria (often requiring that they score highly on a particular measure), they could include proportionally more extreme scores on average than effectiveness studies, which are less selective with their participants as they record changes occurring under ‘naturalistic’ therapy conditions. In clinical research, regression to the mean can create the false impression that very unwell patients have improved more than they actually have, therefore Hsu’s observation that the JT method does not take into account regression to the mean is important to consider. Based on Hsu’s argument, it would make sense to vary from the JT method when participants are selected on the basis of high scores on one measure.

The Gulliksen-Lord-Novick method (GLN; Hsu, 1989; 1995) was developed in response to problems with regression to the mean and was based on the work of Lord and Novick (1968) and Linn and Slinde (1977). It purports to overcome the issue by

¹ A fifth category, *reliable improvement* is proposed by Tingey, Lambert, Burlingame, and Hansen (1996), describes patients who began treatment in the functional population and have made a reliable change in a positive direction. As the current study consists of hospital inpatients who begin treatment in the dysfunctional range, this category will not be included here.
including estimates of the population means and standard deviations toward which scores are assumed to regress. In a similar vein, the Nunnally-Kotsch method (NK; Nunnally & Kotsch, 1983) attempts to account for regression to the mean, but it does so by estimating true pre-treatment scores based on observed pre-treatment scores. The Edwards-Nunnally method (EN; Speer, 1992) also attempts to minimise the effects of regression to the mean, however it approaches the problem differently, by placing a confidence interval (±2 standard errors of measurement) around an estimate of the true pre-treatment score and evaluating the post-treatment score in relation to this confidence interval.

Unlike previously described methods, the Hageman-Arrindell method (HA; Hageman & Arrindell, 1999) attempts to correct for regression to the mean by modifying Jacobson & Truax’s formula for the calculation of the cut-off between the functional and dysfunctional population. The HA method also incorporates the reliability of the outcome measure at pre- and post-treatment into the reliable change calculation. Another method known as Hierarchical Linear Modelling (HLM; Bryk & Raudenbush, 1992) uses longitudinal data (three or more data points) for use in growth curve modelling to classify a patient’s treatment outcome.

Each published clinical significance method claims to have some advantage over others, such as simplicity, or addressing regression to the mean. But, do the different methods produce rates of recovery that are different enough to warrant the use of one method over another? Past research has attempted to address this question by comparing classification rates made by several clinical significance calculation methods.

Bauer, Lambert, and Nielson (2004) compared the rates of deterioration, no change, improvement, and recovery produced from five clinical significance calculation methods: JT, GLN, EN, HA, and HLM. Statistically significant differences were observed between all classification methods except between the JT and GLN methods.
which had a concordance rate of 96%. The HA method produced the most conservative rates of recovery (11.9%) whereas the EN method produced the most liberal rates of recovery (21.2%). Based on these findings, Bauer et al. recommended the use of the JT method due to its simplicity, and because it was a ‘moderate point’ (18.9%) between the extremes of the HA and EN methods.

Speer and Greenbaum (1995) compared classification rates produced by the JT, EN, GLN, NK, and HLM methods and found a high agreement between the methods (77.7%-81.2%), with the exception of the GLN method (51%). Subsequently, it became apparent that the very different classification rates obtained using the GLN method (inflated deterioration rates and very low improvement and recovery rates) were the result of calculation errors made by Speer and Greenbaum (see Hsu; 1999). Recommendations were made to use the HLM method only when identifying deteriorated patients is not a priority, as this method classified no patients as deteriorated. However, such recommendations should be interpreted with caution until research has confirmed their validity.

**Outcome Measures**

When calculating the clinical significance of a patient’s treatment outcome, it is necessary to ask: which outcome measure should be used? In answering this question, one must consider what domain of a patient’s change in functioning is of interest (e.g., general symptom severity, quality of life, depression, anxiety etc.). Classifications made using inappropriate outcome measures can lead to inaccurate conclusions. For example, if a patient’s pre- and post-scores on an anxiety measure classify him or her as ‘improved’, this suggests that the patient has improved in anxiety symptom severity, and must be interpreted as such. Conclusions about improvements in other domains such as quality of life or general symptom severity should not be made based solely on

---

2 The GLN method was labelled as the Hsu-Linn-Lord (HLL) method in Speer and Greenbaum (1995).
this classification. Likewise, meaningful comparisons between patients with similar or
different classifications cannot be made if the outcome measures used differ. Problems
would arise when two equally improving patients were classified as having different
outcomes (e.g., recovered and improved) as a result of two conceptually similar, but
different, outcome measures to classify treatment outcome. For example, consider two
identical patients, A and B, attending two adjacent mental health treatment facilities, 1
and 2, which use different assessment measures to determine patient treatment outcome.

Treatment facility 1 uses a commonly used measure in the public health sector in
Australia, the Kessler Psychological Distress Scale (K10; Andrews & Slade, 2001) and
based on pre- and post-treatment scores on this measure, patient A is classified as

*recovered*. This outcome is considered by treatment facility 1 to indicate a readiness to
terminate therapy, therefore patient A ceases treatment. Meanwhile, patient B attends
treatment facility 2, which uses the Outcome Questionnaire-45.2 (OQ-45.2; Lambert,
Okiishi, Finch, & Johnson, 1998) to assess outcomes, and on the basis of pre- and post-
treatment scores, patient B is classified as *improved*. There are clear possible practical
repercussions of such a scenario; as patient B is not considered to have reached the

*recovered* criteria required for termination, more clinician resources are expended,
perhaps unnecessarily, by continuing this patient in treatment.

Research by Beckstead et al. (2003) compared classification rates made by the
JT method and the Outcome Questionnaire-45.2 (OQ-45.2) against four other outcome
measures: Symptom Checklist 90 Revised (SCL-90R), Social Adjustment Rating Scale
– Self-report/ and Other Report (SAS-SR or OR), Inventory of Interpersonal Problems –
Short Form (IIP-S), and Quality of Life Inventory (QOLI). The average agreement
between the OQ-45.2 and the four outcome measures was 65%. The OQ-45.2 classified
the most (32%) cases as recovered, and the SAS-SR/OR classified the least cases as
recovered (7%). This illustrates that the same sample of patients can be classified quite
differently depending on which outcome measure is used to measure their change from pre- to post-treatment.

The Current Study

To calculate the clinical significance of a treatment outcome, and subsequently, rates of recovery, researchers must choose (a) which of the five published clinical significance calculation methods they wish to use, and (b) which outcome measure is most suitable to use within the chosen calculation. Many possible method-measure combinations are possible; therefore comparisons should be made with caution, while the effects of using different calculation methods and outcome measures are unknown. Previous research has predominantly focused on examining the consistency among classifications produced by (a) different calculation methods when using the same outcome measure (e.g., Bauer, Lambert, & Nielsen, 2004; Speer & Greenbaum, 1995), and (b) different outcome measures when using the same calculation method (e.g., Beckstead et al., 2003). However, no study to date has compared classification rates produced by different outcome measures and different calculation methods simultaneously. The current study aims to address the following questions, using five published clinical significance calculation methods and three popular outcome measures in a large, depressed inpatient psychiatric sample.

To what extent are classification rates consistent when:

1. Different clinical significance calculation methods are used but outcome measures and sample characteristics remain unchanged?

2. Different outcome measures purporting to measure related constructs (i.e., depression, quality of life, general mental health) are used but clinical significance calculation methods and sample characteristics remain unchanged?
**Method**

**Participants**

The sample consisted of 2676 consecutive adult inpatients treated at a private psychiatric facility over a four year period (2004-2008). Patients with an ICD-10 primary diagnosis of a mood disorder from their psychiatrist (78.7% Depressive Episode or Recurrent Depressive Disorder; 19.0% Bipolar Affective Disorder; 0.3% Persistent Mood Disorder) were selected from a pool of patients diagnosed with mood disorders (66%), anxiety disorders (15.6%), substance abuse disorders (7.1%), psychotic disorders (6.1%) and other disorders (3.7%). The sample was 75.9% female, with mean age 42.5 years (SD = 15.3). The mean length of stay was 15.6 days (SD = 10.4). Forty-nine percent were married, 33% were single, and 18% were separated, divorced, or widowed. Data were collected as part of an ongoing evaluation program at the hospital, and written informed consent was obtained at admission. Pre-treatment scores were collected via pen and paper questionnaires on the day of patient’s admission. These scores did not influence the decision to admit the patient to hospital. Similarly, post-treatment scores were collected on the day of patient’s discharge from hospital, and did not influence the decision to discharge the patient. During their stay, patients were involved in a range of group treatments including cognitive behaviour therapy and interpersonal psychotherapy, while receiving ongoing psychiatric care.

**Measures**

Depressive symptomatology was measured using three patient self-report outcome measures; the Depression, Anxiety, and Stress Scale-21 (DASS-21), the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q), and the Mental Health subscale of the Short Form 14 Health Survey Questionnaire (SF-14).

*Depression, Anxiety, and Stress Scale 21.* The DASS-21 is a shortened version of the 42-item version of the DASS which measures levels of depression, anxiety and
stress in the population (Lovibond & Lovibond, 1995). Each subscale contains seven
items which participants are required to respond to on a 4-point scale: 0 = did not apply
to me at all and 3 = applied to me much, or most of the time. A maximum score of 21
on each subscale indicates elevated depression, anxiety, or stress. The chief dependent
variable from this measure was participants’ score on the Depression scale of the
DASS-21. The DASS depression scale has excellent internal consistency in a depressed
clinical sample (.96; Page, Hooke, & Morrison, 2007) and has sound construct validity
(Lovibond & Lovibond, 1995).

Quality of Life Enjoyment and Satisfaction Questionnaire – Short Form. The Q-
LES-Q (SF) is a 14-item self-report measure of the degree of enjoyment and satisfaction
experienced by respondents across various domains of daily functioning (Endicott, Nee,
Harrison, & Blumenthal, 1993). Participants are required to rate their satisfaction with
various aspects of their life such as “household activities” and “physical health” on a 5-
point scale where 1 = very poor and 5 = very good. A maximum score of 100 indicates
high perceived quality of life. The Q-LES-Q shows high internal consistency and test-
retest reliability, and good construct validity (Ritsner, Kurs, Kostizky, Ponizovsky, &
Modai, 2002).

Short Form 14 Health Survey Questionnaire. The SF-14 is a 14-item self-report
measure assessing health on 8 dimensions, including physical functioning, social
functioning, role limitations (physical problems), role limitations (mental problems),
mental health, vitality, pain, and general health perception (SF-14; Brazier et al., 1992).
The chief dependent variable from this measure was participants’ score on the 5-item
Mental Health subscale, in which respondents are required to rate their mental health
over the past month on a 6-point scale where 0 = all of the time and 5 = none of the
time. A maximum score of 100 on this subscale indicates minimal mental health
impairment. The SF-14 has acceptable internal consistency (> .85; Brazier et al., 1992) and acceptable convergent and discriminant validity (Brazier et al., 1992).

**Clinical Significance Calculation Methods**

Five published clinical significance calculation methods were compared in this study. All methods assume continuous data and rely on pre- and post-treatment scores. Four of the five methods (JT, GLN, EN, & NK) use the cut-off scores proposed by Jacobson & Truax (1991). Jacobson and Truax (1991) proposed three different cut-off scores. Cut-off A marks the point at which the current level of functioning falls two standard deviations from the mean of the dysfunctional population (in the direction of the functional population), and calculation requires data relating to the dysfunctional population only. Cut-off B marks the point at which the current level of functioning falls no more than two standard deviations from the mean of the functional population, and calculation requires data relating to the functional population only. Jacobson and Truax (1991) proposed that cut-off C marks the point at which the probabilities of belonging to the functional and the dysfunctional populations are equal, and calculation requires norm data for both populations, which is often not available. Jacobson and Truax stated that cut-off C was preferable to A and B as it uses data from both the functional and dysfunctional populations, which provides information regarding which distribution an individual’s current level of functioning belongs to.

Later, Hsu (1996) criticised cut-off A and B, stating that neither provides enough information to make a confident diagnostic decision. Hsu also demonstrated mathematically that cut-off C did not represent the point at which an individual had an equal probability of belonging to both the functional and dysfunctional populations, but in fact the point at which the proportion of “true members of the functional population” who are misclassified as dysfunctional, is equal to the proportion of “true members of the dysfunctional population” who are classified as functional, when the entire
population is classified. Taking into consideration Hsu’s clarification of cut-off C and his criticism of cut-off A and B, and Jacobson and Truax’s original preference for cut-off C, the current study uses cut-off C in all clinical significance calculations. The Hageman-Arrindell method determines whether a score has passed a cut-off point by calculating a cut-off index, rather than a cut-off score as in the previously described methods. This proposes to overcome the problem of regression to the mean.

**Results**

**Sample Selection**

The original sample consisted of 2676 depressed inpatients for whom pre- and post-treatment questionnaire data were available for the DASS-21. Those who scored 10 or higher (at least “mild” severity) on the depression scale of the DASS-21 at admission were included in the sample (n=2532; 94.6%). For the purposes of the current analysis, only patients who completed all three questionnaire measures at both admission and discharge can be included. A missing values analysis on the initial data set revealed that missing cases were missing completely at random (MCAR; i.e., there is no systematic reason for the missingness); Little’s MCAR test \( \chi^2(45) = 56.15, p=.123 \). When missing cases are missing completely at random, individual cases can be deleted listwise (Enders, 2010). Therefore, patients who did not have SF-14 admission data (n=32) or discharge data (n=35) were removed, leaving 2473 patients. Finally, patients who did not have Q-LES-Q admission data (n=137) or discharge data (n=234) were removed from the sample, leaving a final sample of 2232 (83.4% of original sample).

**Concordance between Outcome Measures**

The correlations between each outcome measure at pre-treatment and post-treatment are shown in Table 1. Strong correlations existed between all three outcome measures at post-treatment.
Table 1

Correlations between Pre-Treatment and Post-Treatment Scores on the Depression Scale of the DASS-21, the Q-LES-Q, and the Mental Health Subscale of the SF-14.

<table>
<thead>
<tr>
<th></th>
<th>Post DEP</th>
<th>Pre MH</th>
<th>Post MH</th>
<th>Pre QOL</th>
<th>Post QOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre DEP</td>
<td>.294**</td>
<td>-.455**</td>
<td>-.208**</td>
<td>-.459**</td>
<td>-.195**</td>
</tr>
<tr>
<td>Post DEP</td>
<td>-</td>
<td>-.190**</td>
<td>-.761**</td>
<td>-.199**</td>
<td>-.664**</td>
</tr>
<tr>
<td>Pre MH</td>
<td>-</td>
<td>.259**</td>
<td>.410**</td>
<td>.228**</td>
<td></td>
</tr>
<tr>
<td>Post MH</td>
<td>-</td>
<td>.198**</td>
<td>.740**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre QOL</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. **p<.01. DEP = Depression scale of DASS-21; MH = Mental Health subscale of SF-14; QOL = Q-LES-Q; Pre = pre-treatment score; Post = post-treatment score.

Correlations between change scores for all measures were significant (p<.01); \( r_{(\text{DEP change} & \text{MH change})} = .633; r_{(\text{DEP change} & \text{QOL change})} = .585; r_{(\text{MH change} & \text{QOL change})} = .582. \)

Criteria for Clinical Significance

Using patient data and norms from outcome measures, cut-off scores were calculated for each clinical significance calculation method (with the NK, EN, and GLN methods sharing their cut-off with the JT method) and each outcome measure (see Table 2). Internal consistency (Cronbach’s alpha), as reported in relevant outcome measure evaluation studies, has been used in all clinical significance calculations.
### Table 2

*Information Required to Calculate Clinical Significance in Three Outcome Measures.*

<table>
<thead>
<tr>
<th></th>
<th>Depression scale of DASS</th>
<th>Q-LES-Q</th>
<th>Mental health subscale of SF-14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-treatment score; M (SD)</td>
<td>35.05 (8.35)</td>
<td>28.14</td>
<td>30.90 (14.34)</td>
</tr>
<tr>
<td>Post-treatment score; M (SD)</td>
<td>17.65 (12.78)</td>
<td>56.16</td>
<td>53.90 (18.77)</td>
</tr>
<tr>
<td>Change score; M (SD)</td>
<td>17.40 (13.05)</td>
<td>28.01</td>
<td>23.0 (19.28)</td>
</tr>
<tr>
<td>Functional population; M (SD)</td>
<td>6.34 (6.97)</td>
<td>82.0 (10.14)</td>
<td>76.0 (17.0)</td>
</tr>
<tr>
<td>Dysfunctional population; M (SD)</td>
<td>30.20 (10.90)</td>
<td>60.0 (14.50)</td>
<td>30.0 (20.0)</td>
</tr>
<tr>
<td>Standard error of difference</td>
<td>3.08</td>
<td>5.12</td>
<td>10.14</td>
</tr>
<tr>
<td>Internal consistencyf</td>
<td>.96d</td>
<td>.95b</td>
<td>.95g</td>
</tr>
<tr>
<td>Cut-off score ‘C’</td>
<td>16.88</td>
<td>62.96</td>
<td>54.54</td>
</tr>
</tbody>
</table>


**Classification Rates**

The first aim of this study was to explore the differences in classification rates made by clinical significance classification methods using three commonly used outcome measures. Figure 1 shows the percentages of patients classified within each
category (recovered, improved, unchanged, or deteriorated) by the five classification methods and three outcome measures.

Figure 1. Proportion of depressed patients classified as recovered, improved, unchanged or deteriorated calculated using combinations of three measures and five clinical significance calculation methods. Note. DEP = Depression Scale of the Depression, Anxiety, and Stress Scale; MH = Mental Health Subscale of the Short-Form Health Survey (SF-14); QOL = Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q); JT = Jacobson & Truax method; NK = Nunnally-Kotsch method; GLN = Gullikson, Lord, & Novick method; EN = Edwards-Nunnally method, HA = Hageman-Arrindell method.

Recovery rates varied widely across classification methods and outcome measures, ranging from 55.1% (calculated using the EN method with scores on the Depression scale of the DASS), to 20.0% (calculated using the HA method with scores on the Q-LES-Q). Deterioration rates ranged from 1.3% (calculated using the HA
method with scores on the MH subscale of the SF-14), to 5.9% (calculated using the NK method with scores on the Q-LES-Q).

Figure 1 demonstrates firstly that there are strong similarities between the rates of patient classifications made using the JT, NK, GLN, and EN classification methods within each measure. The HA method is consistently more conservative in classifying patients as recovered than the other four methods.

Secondly, the three outcome measures differ in terms of their classification rates, with recovery rates at their highest on average when calculated using scores on the Depression scale of the DASS and recovery rates at their lowest on average when calculated using scores on the Q-LES-Q. It is expected that DASS-21 scores will show greater regression to the mean than the SF-14 and Q-LES-Q (i.e., larger improvements, leading to higher recovery rates), as only those patients who at least scored in the mild range for depression according to their DASS-21 scores were included in the analysis.

Thirdly, rates of deterioration observed in the current study (mean 3.0%) are lower than those observed in previous studies that have used outpatient samples (e.g., 6.8% in Speer & Greenbaum, 1995; 8.8% in Bauer et al., 2004). This may reflect the reduced opportunity for patients in a severely dysfunctional inpatient population (in which many patients score at “ceiling”) to worsen enough to be classified as reliably deteriorated.

**Agreement Across Calculation Methods and Outcome Measures**

Table 3 presents agreement between clinical significance calculations across the three outcome measures. Percentage agreement is shown above the diagonal; Kappa agreement statistics are shown below the diagonal. Figures in bold indicate agreement values calculated between the same outcome measure.
Table 3

Agreement Between Clinical Significance Classifications Across the Depression Scale of the DASS-21 (DEP), the Mental Health Subscale of the SF-14 (MH), and the Quality of Life Enjoyment and Satisfaction Questionnaire (QOL).

<table>
<thead>
<tr>
<th></th>
<th>DASS-21 Depression (DEP)</th>
<th>Mental Health (MH)</th>
<th>Quality of Life (QOL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JT</td>
<td>NK</td>
<td>GLN</td>
</tr>
<tr>
<td>JT</td>
<td>-</td>
<td>94.8</td>
<td>98.1</td>
</tr>
<tr>
<td>NK</td>
<td>.92</td>
<td>-</td>
<td>95.0</td>
</tr>
<tr>
<td>DEP</td>
<td>GLN</td>
<td>.97</td>
<td>.92</td>
</tr>
<tr>
<td>EN</td>
<td>.91</td>
<td>.89</td>
<td>.95</td>
</tr>
<tr>
<td>HA</td>
<td>.80</td>
<td>.75</td>
<td>.78</td>
</tr>
<tr>
<td>JT</td>
<td>.43</td>
<td>.40</td>
<td>.42</td>
</tr>
<tr>
<td>NK</td>
<td>.37</td>
<td>.37</td>
<td>.37</td>
</tr>
<tr>
<td>MH</td>
<td>GLN</td>
<td>.40</td>
<td>.38</td>
</tr>
<tr>
<td>EN</td>
<td>.44</td>
<td>.42</td>
<td>.44</td>
</tr>
<tr>
<td>HA</td>
<td>.38</td>
<td>.36</td>
<td>.37</td>
</tr>
<tr>
<td>JT</td>
<td>.30</td>
<td>.28</td>
<td>.30</td>
</tr>
<tr>
<td>NK</td>
<td>.28</td>
<td>.27</td>
<td>.28</td>
</tr>
<tr>
<td>QOL</td>
<td>GLN</td>
<td>.30</td>
<td>.29</td>
</tr>
<tr>
<td>EN</td>
<td>.29</td>
<td>.28</td>
<td>.29</td>
</tr>
<tr>
<td>HA</td>
<td>.24</td>
<td>.22</td>
<td>.23</td>
</tr>
</tbody>
</table>

Note. Percentage agreement is shown above the diagonal; Kappa (κ) is shown below the diagonal. Bold figures indicate comparisons made using the same measure. JT = Jacobson & Truax method; NK = Nunnally-Kotsch method; GLN = Gullikson, Lord, & Novick method; EN = Edwards-Nunnally method, HA = Hageman-Arrindell method.
Agreement is high between classifications made using the same outcome measure (figures in bold). Among these values, the highest levels of agreement occur between the JT, GLN, NK, and EN methods (mean 93.7%; κ = .90). The HA method is the most different from other methods when using the same outcome measure (mean agreement with other methods 81.3%; κ = .72). The highest agreement between outcome measures is between the Depression scale of the DASS-21 and the Mental Health subscale of the SF-14, with an average agreement of 60.3% (κ = .39). The lowest agreement between outcome measures is between the Mental Health subscale of the SF-14 and the Q-LES-Q, with an average agreement of 50.4% (κ = .30).

Overall, the mean agreement between calculation methods was 88.7% (κ = .83), and the average agreement across outcome measures was 53.7% (κ = .33). This suggests that there are greater similarities between classifications made by different calculation methods than there are between classifications made by different outcome measures.

**Discussion**

The current study explored the extent to which inconsistencies exist in classifying a patient’s outcome using different clinical significance calculation methods and outcome measures. The present findings suggest that the choice of outcome measure to be used in a clinical significance calculation method has a greater influence over a patient’s treatment classification than the clinical significance calculation method used. This is certainly not to say that the calculation procedures are not important, but simply that differences in outcome measures account for a larger proportion of the variance in classification rates than differences in calculation procedures.

Within the same outcome measures, results indicated that four of the five published methods for calculating a patient’s outcome classification (i.e., the JT, NK, EN, and GLN methods) were similar; mean agreement was 93.7%. The largest difference observed between these recovery rates within the same outcome measure was
4.4%, between NK and EN in the Mental Health subscale. The similarity is likely due to the shared cut-off point that these four methods use to determine whether a patient has moved from the dysfunctional distribution to the functional distribution, the criteria for achieving a status of recovered. Among these four calculation methods, rates of recovery remained similar, while rates of improvement, no change, and deterioration clearly differed. This is attributable to the components that each classification relies upon. While each of the four classification categories relies upon the reliable change index, the category of recovery is the only one which also relies upon the cut-off score.

Interestingly, while Hsu (1989) argued that the JT method did not address regression to the mean, and the GLN, EN, and NK methods did, there was little difference between the agreement rates that compared the JT method to the NK, GLN, and EN methods (mean agreement within DASS-21 Depression scale was 95.9%), and those that compared the NK, GLN, and EN methods with each other (mean agreement within DASS-21 depression scale was 94.9%). This similarity may suggest that the JT method addressed regression to the mean with similar success to the methods that purport to address it, or that regression to the mean was not present to a great degree. Regression to the mean is more likely when patients are selected based on their high score on a measure, rather than excluded based on their low scores as was the case in the current study.

The Hageman-Arrindell (HA) method was least in agreement with the other four methods (mean agreement 81.3%) and classified markedly fewer patients as recovered than did the four alternative methods. The largest difference between recovery rates calculated using the HA method and the alternatives was 18.4% (calculated between the HA and EN methods within the MH subscale). It was initially predicted that this difference was attributable to the unique cut-off index used by the HA method, however when the cut-off index was used to calculate the cut-off between the functional and
dysfunctional populations, it became clear that these cut-off scores are consistently less conservative than that cut-off score used by the JT method. That is, less change is necessary in order to pass the HA cut-off score. This suggests that the lower rates of recovery calculated when using the HA method are due to differences in the way that reliable change is calculated, rather than differences in the cut-off scores used.

Despite the differences between the HA method and remaining methods, the current findings do not allow any recommendations to be made regarding which clinical significance calculation method should be used to classify patient outcomes. Hsu (1999) noted that it was unwise to make such recommendations based on factors such as higher recovery rates and lower deterioration rates. It may be tempting to use those calculation methods that make more clinically desirable classifications than others (e.g., classifying more patients as recovered, or fewer patients as deteriorated) however this is unwise until research demonstrates that classifications made by one method are more accurate than those made by another. Therefore, future research would do better to focus on the validity of the categorisation rather than the reliability of the methods of calculating clinical significance. For example, if a classification of deterioration by one clinical significance calculation method was a greater predictor of readmission to hospital than another calculation method, then this would provide support for this calculation method over another, and hence it should be recommended for future use. Until such a time when this information is available, it is recommended that authors clearly report the calculation method used, its calculation formula, cut-off scores, as well as means and standard deviations of admission and discharge scores.

In addition, it is important to include the data and source for normative samples, dysfunctional samples, and reliability estimates. The functional distribution can be represented using such samples as asymptomatic populations (i.e., individuals who do not report any symptoms), community samples (i.e., samples that are assumed to
contain some individuals who report symptoms), or the normative sample described in
the development manual for the outcome measure of interest. Regardless of which is
used to represent the functional distribution, it is important that the source of the sample
is clearly reported, to allow appropriate conclusions to be drawn, and comparisons to be
made. Choosing a sample to use to represent the functional distribution should depend
on the purpose for which the functional distribution is required. The source of
parameters (i.e., means and standard deviations) for the dysfunctional sample used
should also be reported. These are typically taken from the study sample, but can also
be taken from the existing literature if the sample characteristics are considered
representative of the study sample. The reliability estimate used should be one of
internal consistency rather than test-retest reliability (see Martinovich, Saunders, &
Howard, 1996; Tingey, Lambert, Burlingame, & Hansen, 1996) and its source should be
clearly reported, whether it has been taken from existing literature or from the current
study sample. A clear consideration of these issues will allow accurate comparisons
between classification rates reported across studies, regardless of which clinical
significance calculation method is used.

In contrast to the overall similarity across calculation methods, large differences
existed between classification rates when different outcome measures were used to
calculate them. The largest difference in recovery rates was 35.1%, observed between
recovery rates calculated using the HA method with the QOL measure (20% classified
as recovered), and the EN method with the DASS-21 measure (55.1% classified as
recovered). This is consistent with previous research by Beckstead et al. (2003) in
which recovery rates yielded from different outcome measures were found to differ
significantly. Indeed, the same has been found in many research studies (Prusoff,
Klerman, & Paykel, 1972; Smith & Glass, 1977; Lambert, Hatch, Kingston, & Edwards,
1986).
Of course, this finding is not surprising given that each outcome measure used in the current study proposes to measure a different psychological construct (depressive symptomatology, general mental health, and quality of life). The existence of different outcome measures means it is unlikely that any given patient will move with the same speed and direction toward the functional distribution of one measure, as he or she will for another measure. The interpretation and relevance of the current findings will vary depending on the context. For example, using society-relevant outcome variables for dependent measures (such as quality of life measures) comes with the high cost of less specificity; a finding which may be relevant in an applied outcomes context. In clinical practice it is obvious that a clinical dependent variable such as a depression measure will change to a larger extent than an indirect indicator such as a quality of life measure in depressed patients who are specifically being treated for their depressive symptoms.

The current study is intended to be a cautionary tale for clinicians; the danger being that clinicians may not consider such differences in outcome measures when patient data is collated and interpreted. But, what impact could ignoring such warnings have on clinical practice? Some may argue that clinical treatment should not be guided entirely by clinical significance classifications, but if clinicians rely on clinical judgment alone there is the danger that incorrect decisions will be made. Hannan et al. (2005) demonstrated that clinicians tended to overpredict patient improvement and were poor at predicting which patients would deteriorate during treatment. More generally, a meta-analysis conducted by Grove, Zald, Lebow, Snitz, and Nelson (2000) showed that “mechanical” predictions such as statistical analyses were generally as accurate as or more accurate than clinical predictions. Therefore it does not come as a surprise that there are some circumstances where clinical significance classifications are used to guide clinical decision making. For example, the Outcomes Questionnaire-45 (OQ-45) is used to provide feedback to clinicians after the completion of each session using a
colour-coded alert system (Lambert, Gregersen, & Burlingame, 2004). When a client is functioning in the normal range according to their OQ-45 responses (white category), clinicians are advised to “consider termination”. When a client’s rate of change is adequate (green category), it is recommended that clinicians stay with their current treatment plan, and when a client’s rate of change is less than adequate (yellow category) it is recommended that treatment is intensified, or a shift in intervention strategies is required. When a client is not making progress as expected (red category) this provides a warning to clinicians that the client may drop out of treatment or have a negative treatment outcome. Although these recommendations are specific to the OQ-45, many clinicians use other outcome measures to complement their own clinical judgment. But, if an outcome measure is slow to detect change or underestimates the magnitude of true change, patients could remain in treatment longer than required. Alternatively, treatment could be prematurely terminated if an outcome measure overestimates the magnitude of true change. It is reasonable to assume that when inaccurate decisions are made regarding the planned duration of patient treatment, it can be detrimental to patients’ well-being as well as financially costly to the healthcare system. While the use of feedback on the OQ-45 leads to improved patient outcomes (Lambert & Shimokawa, 2011), the current study does sound a warning if the data are interpreted without reference to the expected likelihood of improvement for a given patient.

When clinical significance classifications are used to complement clinical judgment, outcome measure selection is important, especially given the variation demonstrated here. This study was not an attempt to compare different outcome measures, to see which would show the largest improvement over the course of a patient's hospital stay. Rather, the aim was to demonstrate more broadly that different outcome measures yield different clinical significance classifications and that this needs
to be considered in situations where such classifications are used to make clinical decisions.

A meta-analysis, such as that conducted by Ogles, Lambert, Weight, and Payne (1990), could be useful in future research to determine the different clinical significance profiles of outcome measures. Ogles and colleagues’ meta-analysis compared 106 agoraphobia outcome studies in which 98 different outcome measures were used to assess constructs related to agoraphobia. Mean effect sizes for outcome measures varied from 2.66 (Phobic Anxiety and Avoidance Scale) to .44 (Heart rate), highlighting the wide range of statistical change reported by different outcome measures used to assess related constructs. Research that compares scales will also permit researchers to identify the extent to which the different clinical significance classification profiles are a function of differences in effect sizes or the thresholds. Another possible direction for future research could be to use longitudinal sampling to identify the individual threshold for endorsement of particular items in a scale. Once the same individuals have been measured on a number of different occasions then analyses derived from Item Response Theory could be used to compare the thresholds of item endorsement and understand how these change over time. Clinicians and researchers will benefit from a richer understanding of the performance across time of the scales used to measure patient change.

As well as calibrating individual outcome measures against each other, the patterns of the difference observed between simultaneously administered outcome measures should also be explored in future research. For example, perceived quality of life may not always occur parallel to symptomatic change, but in some cases, orthogonal to it. A patient who begins treatment severely depressed with a low perceived quality of life may be classified as “recovered” on depressive symptoms post-treatment because of commencement of antidepressant medication which has assisted in
reducing his or her depressive symptoms. However, the patient may remain “unchanged” on perceived quality of life as they high-stress environment in which he or she lives has not changed. Conversely, a patient may report high perceived quality of life owing to factors such as family support despite the persistence of moderate depressive symptoms. Individual “profiles” of differences across combinations of outcome measures could be developed, providing clinically meaningful information to help clinicians target their interventions more specifically to a patient’s individual needs. Clinical validity was not a focus of the current study, however future research should identify variables that relate to varying patient “profiles” of clinical significance classifications made by different outcome measures.

**Recommendations**

It is recommended that future reporting of recovery rates follow particular guidelines in order to ensure that conclusions are drawn in the appropriate context. Firstly, researchers should report the problem (e.g., diagnosis/diagnoses) of the patient(s) classified, to provide support for the choice of outcome measure selected. If rates of recovery are being reported for groups of patients, the percentages of diagnoses must be reported. Next, the classification of the patient(s) should be reported, along with the clinical significance calculation method used and the specific statistics required to replicate the calculation. Then, the particular outcome measure(s) used to evaluate a patient’s outcome needs to be reported, as well as a description of what the outcome measure purports to measure. Finally, the data and source for normative samples and reliability estimates needs to be reported. For example:

“Client A, who presented with major depressive disorder, was classified as recovered as calculated by the Jacobson-Truax method, on the Beck Depression Inventory which purports to measure depressive symptomatology. The normative sample consisted of n participants who did not meet the criteria for major
depressive disorder. Reliability estimates (Cronbach’s alpha) were taken from Study Author (Year).”

Such a statement gives the reader a clear context in which changes have been made and measured, and allows appropriate conclusions to be drawn, and comparisons to be made. In this example, it can be concluded that the patient was assessed on an appropriate measure, given her diagnosis, and that the patient not only made a reliable change in the positive direction, but also passed the cut-off between the dysfunctional and functional population as measured by their scores on the Beck Depression Inventory.

When patient outcomes for a group of patients (e.g., in a hospital clinic) are reported, rates of patient diagnoses need to be included, as well as percentage rates of recovery, improvement, no change, and deterioration. Consider two clinics, both of which treat patients with diagnoses of either major depressive disorder (70%) or generalised anxiety disorder (30%). Clinic A reported that 40% of patients were considered to be recovered at termination whereas Clinic B reported that 60% of patients were considered to be recovered at termination. When outcomes are reported in a comprehensive way, the difference in outcome measure is apparent:

“Clinic A, which consisted of patients with diagnoses of major depressive disorder (70%) and generalised anxiety disorder (30%), was classified as recovered (40%), improved (30%), unchanged (20%), and deteriorated (10%) at termination, as calculated by the Jacobson & Truax method, on the Global Severity Index (GSI) of the Symptoms Check-List 90-Revised (SCL-90R) which purports to measure global symptom severity.”

“Clinic B, which consisted of patients with diagnoses of major depressive disorder (70%) and generalised anxiety disorder (30%), was classified as recovered (60%), improved (20%), unchanged (15%), and deteriorated (5%) at
termination, as calculated by the Jacobson & Truax method, on the Outcome Questionnaire-45 (OQ-45) which purports to measure symptom severity and distress.”

As evident above, the two hypothetical clinics used different outcome measures when calculating patient outcomes. Differences in recovery rates may occur for a multitude of reasons including unique patient, therapist, and treatment variables. As demonstrated in the current study, the choice of clinical significance calculation method and outcome measure can affect recovery rates substantially. Therefore, clear detail regarding each classification should be clearly reported as described above to allow for accurate interpretations and comparisons.
References


Chapter Three

Chapter Three consists of a foreword followed by a manuscript in preparation. A version of this chapter is being submitted as an invited manuscript for a special issue of *Australian Psychologist*. 
As demonstrated in the preceding chapter, different clinical significance classification methods yield different rates of recovery, improvement, no change, and deterioration, but these differences are more pronounced when different outcome measures are used. The question remains: what is the most accurate way to assess the clinical significance of patient change during treatment? The validity of classifications of recovered needs to be assessed by exploring how recovered patients perform on other measures of recovery.

But, how do we best capture the concept of recovery? Surely this depends upon factors that are most important to the particular patient being assessed. For example, if patient A has completed psychological treatment to address her communication and relationship issues, then perhaps it is best to assess this patients’ progress using a measure that assesses the quality of her relationships. Compare this to patient B who has come to therapy to address his difficulties with anxiety in public places. This patient’s change may best be measured using an anxiety symptom measure. It is obviously problematic to compare changes on different measures for different patients and would be time-consuming to assess each patient on different measures depending on their specific treatment goals.

However, if we are interested in each patient’s therapy outcome, why not just ask him or her outright: “how do you feel now?” This question asked post-discharge represents a retrospective rating, meaning that it relies on patient memory. This is in contrast to prospective ratings, which require patients to be followed and have symptoms measured at different time points. If asked a broad question related to changes during treatment, would patients consider the question in regards to their particular issues and answer accurately? Would patients’ perception of their change be valid, and perhaps allow for a determination of whether a change has been clinically
significant? As stated by Andresen, Caputi, and Oades (2010), an appraisal of change from the patients’ perspective is important to include when capturing the concept of recovery. But could it suffice in place of other comparatively complex measures?

To provide an example of how these questions may be answered, we examined them by using one widely used outcome measure; the Depression Anxiety Stress Scales (DASS-21; Lovibond & Lovibond, 1996). When we assess the clinical significance of patient change from pre-treatment to six weeks post-treatment (n=119), using a symptom measure such as the DASS-21, we yield classification rates of: 49% recovered, 12% improved, 34% unchanged, and 5% deteriorated. To compare these clinical significance classification rates to patients’ own perception of their change we asked patients a single question: “how do you feel now compared to how you felt when you arrived at Perth Clinic?” Five response options were possible: “I feel a lot better”, “I feel a bit better”, “I feel the same”, “I feel a bit worse”, and “I feel a lot worse”. Five response options are provided so as to create a symmetrical rating system, with the most neutral response (“I feel the same”) in the middle of the scale. The same sample of patients answered this question six weeks following their discharge from hospital, to allow comparability with clinical significance classifications. Of these patients, 67.2% reported “I feel a lot better”; 20.2% reported “I feel a bit better”; 3.4% reported “I feel the same”; 5.0% reported “I feel a little worse”; and 4.2% reported “I feel a lot worse”. Patient perception of change was non-normally distributed, with skewness of -2.02 (SE=.22) and kurtosis of 3.30 (SE=.44). Therefore when patients are asked to reflect on their change at a single time period, the majority report a positive one.

When these ratings are matched to what clinical significance classifications purport to represent qualitatively, a rating of feeling “a lot better” aligns with recovered, feeling “a bit better” aligns with improved, feeling “the same” aligns with unchanged, and feeling either “a bit worse” or “a lot worse” align with deteriorated. Figure 1 shows
the classification rates resulting when 119 patients’ outcomes are evaluated in both ways: clinical significance classification and self-perception of change.

Figure 1. Classifications resulting from 119 patients assessed using Jacobson-Truax (1991) clinical significance classification method based on scores on the DASS-21 and assessed using a self-perception of change measure.

Only 13% of patients reported either feeling the “same” or feeling “worse”. When compared to the clinical significance classifications made based on symptom measures taken at the same time of self-perception ratings, the results are striking, with 39% of patients considered to have either made no reliable change or reliably deteriorated.

Clearly, outcome classifications based on the clinical significance of changes on a symptom measure contrast with what patients report through response to a single item six weeks after their discharge. The current finding supports past literature exploring the biases present in retrospective versus prospective ratings of symptoms. Research shows that patients tend to underreport past symptoms when assessed retrospectively (Simon
Patients who have experienced shorter illness duration tend to be more unreliable in their retrospective ratings (Foley, Meale, & Kendler, 1998). Furthermore, when patients are experiencing symptoms at the time of the retrospective rating of symptoms, their recall of past symptoms increased, especially with increasing symptom severity and chronicity (Wells & Horwood, 2004). More recently, Moffitt and colleagues (2010) showed that prospective estimates of symptom presence and severity exceeded retrospective estimates by double. These findings together demonstrate the unreliability of retrospective ratings of symptoms, potentially due simple forgetfulness, or other reporting biases.

It could be that patients are responding in a socially desirable way, in wanting to portray to the researcher a greater change than had been experienced. Additionally, when patients answer a single item following treatment, they are relying on the memory of their symptom severity at admission to hospital. Current judgments are biased by current mood state (Bower, 1981; Johnson & Magaro, 1987) therefore assessing patient change using a single self-report item completed at one time period is unwise. These subjectivities could be overcome by creating a benchmark for determining the success of a treatment by combining judgments from relevant stakeholders (e.g., clinicians, patient's family and/or close friends). Such a process would still result in a subjective construct, however this would be less impacted by patient mood bias, and therefore increase the accuracy with which meaningfulness of change could be assessed. This is beyond the scope of the current dissertation but could be considered in future research. Alternative questions such as “did you feel treatment was worthwhile” or “were you satisfied with treatment” may assess a different aspect of the meaningfulness of the change for the patient but could still be influenced by mood bias. By comparison, measuring patients at pre- and post-treatment (i.e., a prospective rating of symptoms)
allows for calculation of a difference score, which reduces potential reporting biases
associated with retrospective rating of symptoms.

Given the findings described here, it is necessary to go beyond simply asking
patients how they feel after treatment when aiming to validate how recovery is
conceptualised by clinical significance classification methods. This will be important in
the following chapter, which aims to examine the construct validity of two clinical
significance calculation methods with respect to the way recovery is evaluated. It is vital
that patients and treatment providers are provided with valid and objective information
regarding patient outcomes when clinical significance methods are used to evaluate
them.
References


Comparing the validity of clinically significant change classifications yielded by the Jacobson-Truax method and the Hageman-Arrindell method

Abstract

Several studies have examined the ecological validity of classifications of clinically significant change, and findings suggest that clinical significance classifications have some ecological validity. However, since rates of recovery vary depending on the calculation method used (Ronk, Hooke, & Page, 2012; Chapter 2), it is important to compare the validity of classifications yielded by each. The current study investigated the extent to which classifications of clinically significant change based on the DASS-21 captured important elements of recovery when two calculation methods were used: the Jacobson-Truax method (Jacobson & Traux, 1991) and the Hageman-Arrindell method (Hageman & Arrindell, 1999). Inpatients (n=355) discharged from a private psychiatric clinic were invited to complete questionnaires six weeks after their discharge. In both the JT and HA methods, clinically significant change on the DASS-21 was related to a greater consumer-based sense of recovery, greater perceived quality of life, and fewer readmissions to hospital within 28 days of discharge. Since there was found to be no advantage to using one method over another when recovery is of interest, the use of the JT method is recommended as it is simpler, and is more commonly used.
Evaluations of patient change are important in assessing treatment effectiveness and guiding quality improvement in mental health (Lambert & Ogles, 2009; Newnham & Page, 2010). Statistical methods of determining clinical significance are intended to provide clinicians with a meaningful classification of patient outcomes based on a selected outcome measure. The most widely used calculation method for clinical significance is the Jacobson-Truax method (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991) which considers the reliability of the change made (Reliable Change Index; RCI) in the context of the distribution that the patient belongs to post-treatment (e.g., functional or dysfunctional). A patient change is considered *clinically significant* if the patient has moved reliably into the functional distribution. This type of movement is also termed *recovered*. Patients have *improved* if they have made a reliable change but remained in the dysfunctional population, *unchanged* if they have not made a reliable change, and *deteriorated* if they have made a reliable change in a negative direction.

Researchers and publishers are increasingly recommending the reporting of rates of clinically significant change (e.g., American Psychological Association); therefore it is important that the classifications clinical significance methodology produces are valid. That is, a patient who is classified as *recovered* should demonstrate aspects of recovery as measured by other variables, also known as convergent validity, or ‘ecological’ validity. Ecological validity refers to the capacity of an instrument to capture critical elements of the real world environment (Schmuckler, 2001). Examples of real world factors related to a classification of recovered in mental health may include having a positive perception of one’s quality of life, remaining out of hospital, and having goals for the future.
Several research studies have examined aspects of the ecological validity of the Jacobson-Truax method for assessing clinical significance. Clinically significant change on the Symptom Check List-90 Revised (SCL-90R; Derogatis, 1977) relates to patient’s satisfaction with their therapy (Ankuta & Abeles, 1993). Similarly, reliable change on symptom measures as assessed by the JT method relates to client-rated and therapist-rated perceived change (Lunnen & Ogles, 1998). The convergent validity of classifications of clinically significant change for depressed patients across three different depression measures (Beck Depression Inventory; Hopkins Symptom Checklist; and Hamilton Rating Scale for Depression) has also been demonstrated. Classifications of clinically significant change within this group were in agreement for 73% of patients (Ogles, Lambert, & Sawyer, 1995). In 2003, Beckstead and colleagues added further evidence of the convergent validity of clinically significant change by exploring the correlates of clinically significant change on the Outcomes Questionnaire-45 (OQ-45; Lambert, Okiishi, Finch, & Johnson, 1998). They found that those patients who had made a clinically significant change on the OQ-45 were likely to have also made clinically significant change on the SCL-90R, the Social Adjustment Rating Scale, the Inventory of Interpersonal Problems, and the Quality of Life Inventory (Beckstead et al., 2003). Newnham, Harwood and Page (2007) determined that clinically significant change on the Medical Outcomes Short Form Questionnaire (SF-36; Brazier et al., 1992) was associated with a greater perceived quality of life, as well as greater clinician-rated functioning. Furthermore, Wise (2010) demonstrated that 56% of substance abuse patients who made a reliable change on the SCL-90R had a clinically meaningful change in the percentage of days abstinent from substances, further providing evidence for the validity of classifications of clinical significance. Most recently, Ronk and colleagues (2012; i.e., Chapter 2) demonstrated that when clinical significance based on the JT method is assessed using different measures related
to depression (Quality of Life Enjoyment Scale; Depression scale of the DASS-21; and SF-36 Mental Health Scale), the results are largely convergent, with differences as expected, given that symptoms of depression and quality of life are conceptually different. Therefore, these research findings together support the ecological validity of clinical significance classifications for the JT method of calculating the clinical significance of patient change.

However, there exist several other methods by which to determine the clinical significance of a treatment outcome, including the Gulliksen Lord Novick method (GLN; Hsu, 1989, 1995), the Nunnally-Kotsch method (NK; Nunnally & Kotsch, 1983), the Edwards-Nunnally method (EN; Speer, 1992), the Hageman-Arrindell method (HA; Hageman & Arrindell, 1999) and Hierarchical Linear Modelling (HLM; Bryk & Raudenbush, 1992). Similarly, McGlinchey, Atkins and Jacobson (2002) compared clinical significance classifications made using the JT, GLN, EN, HA, and HLM methods, and found that the HA method classified patients significantly differently to the JT, GLN, and EN methods. They stated that the HA method was less sensitive, since a greater amount of patient change was required in order for a patient to be considered reliably improved. Similarly, Ronk, Hooke, and Page (2012; i.e., Chapter 2) compared rates of clinically significant change of a sample of 2,232 depressed inpatients calculated using five calculation methods (JT, GLN, NK, EN, and HA) and three different outcome measures. While four of the five calculation methods examined yielded similar rates of clinically significant change, the HA method produced consistently distinct classifications. For example, while 53.3% of patients were considered recovered on the Depression scale of the DASS-21 according to the JT method, only 41.1% were considered recovered based on calculations made by the HA method. Similar patterns were observed for the other two outcome measures used to assess patient change. Therefore it appears that the HA method is more conservative in
assigning classifications of \textit{recovered} to patients than other methods, including the popular JT method.

The reason for such differences between the classification rates lies in the method of calculations. The HA method proposed to more closely approximate true scores than the originally proposed JT method. The JT method classifies a patient’s outcome based on the reliability of the pre- to post-treatment change and whether or not the patient has moved from the dysfunctional population to the functional population. The calculation uses patients’ observed scores, which contain measurement error. The HA method attempts to correct regression to the mean by using an approximation of true scores rather than observed scores. In addition, the HA method uses a different method to determine whether a patient belongs to the functional or dysfunctional distribution. While the JT method uses a cut-off score to separate the functional and dysfunctional distributions, the HA method uses a cut-off index score which allows users to determine that a patient has passed the cut-off score with 95% confidence that this is correct.

These differences between rates of recovery based on the JT and HA methods of classifying change cannot be ignored. However, as previously stated by Hsu (1999), one method cannot be recommended over another based purely on higher or lower rates of classifications of clinically significant change. It is important to determine whether one method’s recovered patients experience changes in other areas of importance, such as quality of life, that reasonably correspond with the concept of recovery, when compared to other classification methods. In their study, McGlinchey, Atkins and Jacobson (2002) demonstrated that clinical significance classifications made by the five methods examined were able to discriminate between those patients who relapsed during the subsequent two years following treatment, and those who did not. They suggested that
there is no difference in the accuracy of one method over another in predicting relapse within two years, but they recommended that more research was necessary to determine which method is more “useful” (p544). Similarly, Lambert and Ogles (2009) recommended that research be conducted to ascertain which method is better able to detect “meaningful” (p499) changes in patients who undergo treatment.

Thus, it is necessary to determine which method is more useful or better at capturing meaningful changes in patients before recommending the use of one calculation method over another. In other words, an analysis of the construct validity of each method’s definition of recovery is required. Comparisons between both methods’ recovered patients will be conducted with respect to variables that are important to both clinicians and patients. Several variables will be used in the current study to evaluate how usefully and meaningfully each method captures the construct of ‘recovery’. Analyses will only be conducted using the JT and HA methods, since the remaining three methods explored in McGlinchey et al., (2002) and Ronk et al. (2012; i.e., Chapter 2) were largely similar to the JT method and therefore, minimal differences could be expected.

Recovery Evaluation Variables

It is often implied in clinical research that symptom reduction is related to the concept of recovery, especially when symptom measures (e.g., SCL-90R) are used to assess the clinical significance of patient change during treatment. However, it is important to consider the definition of recovery as being broader than symptom amelioration alone. The concept of ‘consumer-based recovery’ captures the notion that there are many facets to recovery, wider than symptom reduction, including hope, healing, empowerment, self-identity, pursuing meaningful goals, developing connections with others, and having a sense of control (Clarke, Oades, Crowe, Caputi,
& Deane, 2009; Jacobson & Greenley, 2001). This definition of recovery posits that a focus on symptom reduction alone is too narrow, as patients who report severe symptoms can still experience improvements in other aspects of their lives (Davidson, Drake, Schmutte, Dinzeo, & Andres-Hyman, 2009; Zimmerman et al., 2012). Therefore, the first recovery evaluation variable is a consumer-based measure of recovery; the Recovery Assessment Scale (RAS; Corrigan, Salzer, Ralph, Sangster, & Keck, 2004). If clinically significant change on a symptom measure as classified by the JT and HA methods corresponds with higher scores on the RAS, then this will suggest that clinical significance methodology can converge with consumer-based recovery measures, therefore demonstrating ecological validity. The RAS originally had 41 items, however Hancock et al. (2011) removed 10 items due to poor fit statistics or item redundancy, resulting in a 31-item scale. They suggested that the 10 removed items failed to fit the recovery construct, and this is supported by weak factor loadings in previous studies. Hancock et al. argued that the remaining 31 items relate closely to processes associated with consumer based recovery in the literature (e.g., symptom management, a sense of control). In a review of recovery measures (Burgess, Pirkis, Coombs, & Rosen, 2011), the RAS was recommended as a measure of consumer-based recovery as it is acceptable to consumers, is brief and easy to administer, and it has sound psychometric properties. Most recently, Sklar, Groessl, O’Connell, Davidson, and Aarons (2013) conducted a systematic review of mental health recovery measures and highly recommended the use of the RAS due to its strong psychometric properties and high correspondence of its items with definitions of consumer-based recovery.

Another domain associated with the consumer-based conceptualisation of recovery is quality of life. Knowing that a clinically significant change has occurred on a symptom measure is useful information, but it is also important to examine how these changes are reflected in a patients’ quality of life (Gladis, Gosch, Dishuk, & Crits-
Christoph, 1999), since mental illness affects a person’s functioning across several domains. Therefore, the second recovery evaluation measure chosen is the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q; Endicott, Nee, Harrison, & Blumenthal, 1993). This instrument assesses patients’ enjoyment and satisfaction with their lives; arguably an important factor in considering whether a patient has recovered. It has been demonstrated that an increase in symptoms relates to a decrease in quality of life for patients with Major Depressive Disorder (Trivedi et al., 2006). If the inverse of this is also true, and clinically significant change on a symptom measure relates to significant improvements in perceived quality of life, then this will provide convergent validity for classifications of recovered or clinically significant change.

The two measures discussed thus far are subjective, since they rely on patient report. Therefore it is important to include an objective indicator of treatment outcome in our set of recovery evaluation variables. Whether or not a patient has been readmitted to an inpatient facility soon after their discharge has a logical relationship to outcome, and it can be objectively measured using hospital records. Readmission to hospital specifically within the 28 day period following discharge is used as a national clinical indicator of the quality of care received according to the Clinical Indicator Guidelines published by the Australian Council on Healthcare Standards (ACHS, 2012). Therefore this variable will be used as our third recovery evaluation variable. Readmission to hospital within 28 days is considered a poor outcome, and is associated with more severe symptoms (Lyons et al., 1997), a psychotic diagnosis (Hodgson, Lewis, & Boardman, 2001), and prior admissions to hospital (Callaly, Trauer, Hyland, Coombs, & Berk, 2011). Specifically, a person who is readmitted to hospital within 28 days is not consistent with that person being considered to be recovered (assuming the readmission is for the same reason as the index admission). Conversely, a lack of readmission to hospital is consistent with a treatment success. For example, greater patient-reported
symptom improvement during treatment is associated with fewer subsequent readmissions to hospital (Byrne, Hooke, & Page, 2010).

**Outcome Measure Selection**

It is important to select an outcome measure that captures the elements of patient change that are most relevant to the purposes of the outcome classification. The current study will use the Depression Anxiety Stress Scales 21 (DASS-21; Lovibond & Lovibond, 1995). It has been chosen as it is a quick to administer, free, self-report symptom measure commonly used in Australia that assesses symptoms of depression, anxiety, and stress. The majority of the current sample has primary diagnoses of either mood or anxiety disorders (75.5%); therefore the measure is appropriate in regards to assessing changes in symptoms during treatment. In addition, the relevant values for calculating the clinical significance of a change on the DASS-21 are published and the resulting categories demonstrate convergent validity with another symptom measure (the SF-36; see Ronk, Korman, Hooke, & Page, 2013; i.e., Chapter 5). The relevant psychometric properties of the DASS-21 are discussed in the Methods section of this article.

**Aims and Hypotheses**

The focus of the current study is on the categorisation of patients as recovered, which is defined according to statistical methods for determining clinical significance as both (a) making a statistically reliable change during treatment; and (b) belonging to the ‘functional’ population at post-treatment. Firstly, we aim to contribute to the growing body of literature examining the validity of the JT method for assessing clinically significant change by exploring the relationship between classifications of recovered and three variables related to the concept of recovery. It is necessary to explore the links between clinical significance classifications and these criterion measures before any
further assumptions can be made about the validity of the clinical significance methodology. It is hypothesised that those patients who are classified as *recovered* on the DASS-21 by either the JT or HA methods, when compared to those who are not classified as *recovered*, will:

(a) score higher on the 31-item Recovery Assessment Scale (RAS; Corrigan, Giffort, Rashid, Leary, & Okeke, 1999; Hancock et al., 2011) indicative of a greater sense of consumer-based recovery;

(b) score higher on the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q; Endicott et al., 1993), indicative of a greater perceived enjoyment and satisfaction with life; and

(c) have lower rates of readmission to hospital within 28 days of their discharge, indicative of a more successful post-discharge period.

Secondly, if it is found that there is an association between making a clinically significant change and our three criterion measures, we aim to determine whether one method demonstrates more convergent validity than another. That is, is the more conservative method (HA) better ‘calibrated’ to recovery than the more lenient method (JT)? Recovered patients will be compared with non-recovered patients across both the JT and HA calculation methods in relation to the recovery evaluation variables to determine which method best classifies recovered patients.

**Method**

**Participants**

Participants were selected from a sample of 718 consecutively admitted patients with complete data who had been discharged from a private psychiatric hospital in Perth, Western Australia between April 2011 and January 2012. The mean age of patients was 42.9 years (SD=15.1) and the mean length of stay in hospital was 17.4 days.
Married patients made up 50.3% of the sample, 33.2% were single, and 16.5% were separated, divorced, or widowed. Participants were given a primary diagnosis by their treating psychiatrist. The sample consisted of patients with primary diagnoses of mood disorders (56.1%), anxiety disorders (19.4%), substance abuse disorders (13.7%), psychotic disorders (5.9%) and other diagnoses including personality disorders (4.9%).

While in hospital, patients complete a range of group therapies led by clinical psychologists and occupational therapists including cognitive behavioural therapy, interpersonal therapy, and structured activity-based therapy while under the care of nursing staff and their psychiatrist. As part of routine quality assurance at the hospital, patients were invited complete questionnaire measures at both admission and discharge. Participants provided informed consent and the study had ethical approval.

**Measures**

*Recovery Assessment Scale.* Scores on the RAS have strong internal consistency ($\alpha=.93$) and good test-retest reliability ($\alpha=.88$; Corrigan, Giffort, Rashid, Leary, & Okeke, 1999). The validity of RAS score interpretations demonstrates good convergent validity (Corrigan et al., 1999), and this has been tested on an Australian sample (McNaught, Caputi, Oades, & Deane, 2007) with correlations with other recovery-oriented scales ranging from $r=.20 - .68$. The RAS demonstrates divergent validity from symptom or function-based measures such as the Health of the Nation Outcome Scales (HoNOS; Wing et al., 1998), which suggests that the RAS assesses recovery in a broader sense than symptoms and functionality alone.

Based on results of the Rasch analysis conducted by Hancock et al. (2011), the 31-item version of the RAS with a five-point rating scale (1 = strongly disagree; 5 = strongly agree) was chosen to determine recovery scores in the current study. Scores on
each item are summed to form one score representing ‘recovery’, with a minimum possible score of 31 and a maximum possible score of 155. A higher score indicates a stronger experience of recovery in a broader sense than symptom reduction alone.

*Depression Anxiety Stress Scales 21 (DASS-21).* The DASS-21 measures levels of depression, anxiety, and stress. Respondents rate 21 items such as “I felt down-hearted and blue” and “I felt that life was meaningless” on a scale ranging from zero (did not apply to me at all) to three (applied to me very much, or most of the time). Within each scale, the total score is doubled so that the minimum score is zero and the maximum score is 42. The scores on each scale have adequate internal consistency (α=.88 for Depression; α=.82 for Anxiety and α=.90 for Stress; Henry & Crawford, 2005) and the interpretations of the construct demonstrate good convergent and discriminant validity (Henry & Crawford, 2005; Page, Hooke, & Morrison, 2007).

*Quality of Life Enjoyment and Satisfaction Questionnaire- Short Form (Q-LES-Q; Endicott et al., 1993).* The Q-LES-Q is a 14 item self-report scale assessing quality of life across domains such as physical health, and household activities. Respondents rate their satisfaction with each domain on a 5-point scale ranging from 1 (very poor) to 5 (very good). Item scores are added and transformed onto a scale ranging from minimum of 0 to maximum of 100, with higher scores indicative of higher perceived quality of life. Scores on the Q-LES-Q demonstrate high internal consistency (>.90) and test-retest reliability (.63-.89), and the interpretations of the scores show good construct validity (Ritsner, Kurs, Kostizky, Ponizovsky, & Modai, 2002; Trivedi et al., 2006).

*Readmission.* Readmission to hospital within 28 days of discharge was assessed using hospital admission records.

**Clinical Significance Calculation Methods**

The current study used the Jacobson-Truax method of clinical significance classification (JT; Jacobson et al., 1984; Jacobson & Truax, 1991) and the Hageman
Arrindell method of clinical significance classification (Hageman & Arrindell, 1999). Jacobson and Truax aimed to classify patients based on the reliability and magnitude of the change they had made. Classifications are made on the basis of two criteria: (1) whether the patient has passed a cut-off point between the functional and dysfunctional population, and (2) whether the magnitude of the change is statistically reliable. The cut-off between the dysfunctional population and the functional population can be calculated using one of three formulas depending on the availability of valid dysfunctional and functional population norms, and whether these population distributions overlap (for all three formulas see Jacobson et al., (1984); Jacobson & Truax (1991). The present study used cut-off ‘C’ to represent the cut-off between the functional and dysfunctional population as recommended by Hsu (1996), which is calculated as follows:

\[
\text{Cut-off } C = \frac{(SD_0 M_1) + (SD_1 M_0)}{SD_0 + SD_1}
\]

where: 

- \( M_0 \) = mean of the normal/functional population
- \( SD_0 \) = standard deviation of normal/functional population
- \( M_1 \) = mean of the pre-treatment scores/dysfunctional population
- \( SD_1 \) = standard deviation of pre-treatment scores/dysfunctional population

The reliable change index (RCI) is calculated using the following formula:

\[
RCI = \frac{(x_2 - x_1)}{\sqrt{2S_E^2}}
\]

where: 

- \( S_E = SD_1 \sqrt{1 - r} \); (standard error of measurement)
- \( x_1 \) = individual’s pre-treatment score
- \( x_2 \) = individual’s post-treatment score
- \( r \) = reliability of measurement instrument (Cronbach’s alpha)
Positive reliable change is indicated by an RCI score greater than 1.96 when using positive measurement instruments (i.e., lower scores reflect greater dysfunction) and less than -1.96 when using negative measurement instruments (i.e., higher scores reflect greater dysfunction). If a patient makes a positive reliable change and passes the cut-off point then they are classified as recovered. If a patient makes a positive reliable change but remains in the dysfunctional population, they are classified as improved. When no reliable change has been made, a patient is classified as unchanged, and deterioration occurs when a patient has made a negative reliable change.

Hageman and Arrindell (1999) proposed a calculation method which modifies the JT formula for measuring reliable change and for measuring the cut-off between the functional and the dysfunctional population by incorporating the reliability of the outcome measure at pre- and post-treatment. The reliable change index is calculated using the following formula:

\[
RCI = \frac{(x_1 - x_0)r_{DD} + (M_1 - M_0)(1 - r_{DD})}{\sqrt{r_{DD}}} \sqrt{2S_E^2}
\]

where: \(x_0\) = individual’s pre-treatment score
\(x_1\) = individual’s post-treatment score
\(M_0\) = mean of pre-treatment scores
\(M_1\) = mean of post-treatment scores
\(SD_0\) = standard deviation of pre-treatment scores
\(SD_1\) = standard deviation of post-treatment scores
\(S_E\) = standard error of measurement
\(r_{DD}\) = reliability of difference scores; which is calculated by:

\[
r_{DD} = \frac{SD_0^2 r_c + SD_1^2 r_c - 2SD_0 SD_1 r_c}{SD_0^2 + SD_1^2 - 2SD_0 SD_1 r_c}
\]

where: \(r_c\) = correlation between pre-treatment and post-treatment scores
\[ r_0 = \text{reliability of pre-treatment scores} = \frac{SD_0^2 - SE^2}{SD_0^2} \]
\[ r_1 = \text{reliability of post-treatment scores} = \frac{SD_1^2 - SE^2}{SD_1^2} \]

A change is considered reliable if the RCI is greater than 1.65 when using ‘positive’ measurement instruments (i.e., where a higher score indicates greater functioning) and if the RCI is less than -1.65 when using ‘negative’ measuring instruments (i.e., where a higher score indicates greater dysfunction).

In contrast to the JT method, the HA method calculates a cut-off index, rather than a cut-off score. An index score that has an absolute value greater than 1.65 indicates that there is a 95% probability that the classification has been made in the correct direction. The individual is considered to have passed the cut-off for clinical significance if the index score is greater than 1.65 for ‘positive’ measurement instruments, and less than -1.65 for ‘negative’ measurement instruments. The cut-off index for the HA method is calculated using the following formula:

\[
CS_{indiv} = \frac{M_1 + (x_1 - M_1)r_1 - TRC}{\sqrt{r_1}SE}
\]

where: TRC = ‘true’ cut-off score for clinically significant change and is estimated using the following formula:

\[
TRC = \frac{SD_0\sqrt{r_x}M_0 + SD_0\sqrt{r_0}M_x}{SD_x\sqrt{r_x} + S_0\sqrt{r_0}}
\]

where: \( r_x = \frac{SD_x^2 - SE^2}{SD_x^2} \)

\( M_x = \text{mean of normal population} \)

\( SD_x = \text{standard deviation of normal population} \)
In addition to classifications of clinical significance made using scores from the time of pre-treatment to post-treatment, classifications were also calculated using pre-treatment scores and scores at six weeks post-treatment.

**Design and Procedure**

Patients completed the DASS-21 and Q-LES-Q at admission and discharge as part of routine outcome monitoring, which patients give written consent to prior to admission. A total of 718 patients discharged during the period from April 2011 and January 2012 completed both measures at admission and at discharge. These data were used to assess clinical significance of change from pre- to post-treatment as measured by the DASS-21 as well as quality of life score at discharge and readmission to hospital within 28 days of discharge.

A cohort of 355 patients who were discharged during the first half of the study period was invited to complete the DASS-21 and RAS six weeks after their discharge. Only patients who were staying at their current mailing address were contacted. The total response rate was 41.1%, which compares favourably with other mail-out surveys (e.g., Church, 1993; Kaplowitz, Hadlock, & Levine, 2004). Only those cases with complete data (n=119) were used in the analyses of RAS total scores and DASS-21 scores six weeks post-discharge. Age differed significantly between responders ($M=48.0$ years; $SD=15.7$) and non-responders ($M=40.4$ years; $SD=15.0$); $t(353)=4.68$, $p<.05$. In addition, the length of stay in hospital was longer ($M=19.2$ days; $SD=15.8$) for patients who responded compared to patients who did not respond ($M=16.1$ days; $SD=12.2$); $t(353)=2.16$, $p<.05$. There were no significant differences between responders and non-responders in relation to symptom severity at admission or discharge and prior admissions to hospital suggesting that the final sample was generalizable to the complete group on these dimensions.
Independent sample $t$-tests and corresponding measure of effect size, Cohen’s $d$, will be used to evaluate the difference in RAS scores and quality of life scores for those patients who make a clinically significant change on the DASS-21 and those who do not. Chi-squared analysis ($\chi^2$) and corresponding measure of effect size, phi ($\phi$) will be used to assess the difference in readmission rates within 28 days between those who patients who make a clinically significant change on the DASS-21 and those who do not.

**Results**

**Descriptive Statistics**

Mean and standard deviations for each measure at pre-treatment, post-treatment, and six weeks post-treatment are shown in Table 1. Mean scores for each scale of the DASS-21 decreased between pre-treatment and post-treatment, and quality of life increased.
Table 1

Means and Standard Deviations (in Parentheses) for the DASS-21, Q-LES-Q and RAS at Pre-Treatment, Post-Treatment, and Six Weeks Post-Treatment

<table>
<thead>
<tr>
<th></th>
<th>DASS-21 scale</th>
<th>Q-LES-Q</th>
<th>RAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depression</td>
<td>Anxiety</td>
<td>Stress</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-treatment</td>
<td>718</td>
<td>28.27</td>
<td>20.53</td>
</tr>
<tr>
<td></td>
<td>(11.84)</td>
<td>(10.85)</td>
<td>(10.24)</td>
</tr>
<tr>
<td>Post-treatment</td>
<td>718</td>
<td>12.82</td>
<td>10.18</td>
</tr>
<tr>
<td></td>
<td>(10.79)</td>
<td>(8.95)</td>
<td>(9.69)</td>
</tr>
<tr>
<td></td>
<td>(11.69)</td>
<td>(9.16)</td>
<td>(10.53)</td>
</tr>
<tr>
<td>Range of measure</td>
<td></td>
<td>0 – 42*</td>
<td>0 – 42*</td>
</tr>
</tbody>
</table>

Note: * = higher scores reflect more negative functioning. ** = higher scores reflect more positive functioning. DASS-21 = Depression Anxiety Stress Scales – 21. Q-LES-Q = Quality of Life Enjoyment and Satisfaction Questionnaire. RAS = Recovery Assessment Scale.

Correlations between DASS-21 scales at pre-treatment, post-treatment, and between change scores were moderately inter-correlated, as has been shown in the literature. Pre-treatment correlations were significant ($p<.01$); $r_{(Dep \& Anx)} = .52$, $r_{(Dep \& Str)} = .57$, and $r_{(Anx \& Str)} = .69$. Post-treatment correlations were significant ($p<.01$); $r_{(Dep \& Anx)} = .70$, $r_{(Dep \& Str)} = .78$, and $r_{(Anx \& Str)} = .74$. Correlations between change scores were also significant ($p<.01$); $r_{(Dep \& Anx)} = .59$, $r_{(Dep \& Str)} = .68$, and $r_{(Anx \& Str)} = .69$.

Percentages of patients classified into each clinical significance category according to the Jacobson-Truax (1991) method and the Hageman-Arrindell (1999) method are shown in Table 2. As can be seen, when patients are classified using the JT.
method, there are higher rates of clinically significant change than when the HA method is used. This suggests that the HA method is more stringent in its classification of recovery (see also Ronk et al., 2012; i.e., Chapter 2). It is also evident that classifications of deteriorated patients yield identical proportions with both calculation methods.
Table 2

Percentage of Patients Classified into each Clinical Significance Category by the Jacobson-Truax (JT) Method and the Hageman-Arrindell (HA) Method Based on DASS-21 Scale Scores Calculated Across Two Time Periods

<table>
<thead>
<tr>
<th></th>
<th>DASS-21 Depression scale</th>
<th>DASS-21 Anxiety scale</th>
<th>DASS-21 Stress scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JT method</td>
<td>HA method</td>
<td>JT method</td>
</tr>
<tr>
<td>Pre-treatment to post-treatment (n=718)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovered</td>
<td>57.5%</td>
<td>42.1%</td>
<td>36.8%</td>
</tr>
<tr>
<td>Improved</td>
<td>14.9%</td>
<td>35.9%</td>
<td>14.6%</td>
</tr>
<tr>
<td>Unchanged</td>
<td>25.5%</td>
<td>19.9%</td>
<td>46.4%</td>
</tr>
<tr>
<td>Deteriorated</td>
<td>2.1%</td>
<td>2.1%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Pre-treatment to six weeks post-treatment (n=119)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovered</td>
<td>49.5%</td>
<td>37.8%</td>
<td>36.1%</td>
</tr>
<tr>
<td>Improved</td>
<td>11.8%</td>
<td>29.4%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Unchanged</td>
<td>34.5%</td>
<td>28.6%</td>
<td>47.9%</td>
</tr>
<tr>
<td>Deteriorated</td>
<td>4.2%</td>
<td>4.2%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>
Of the 718 patients discharged from hospital between April 2011 and January 2012, 64 (8.9%) were readmitted to hospital within 28 days of their discharge from hospital.

**Recovery Assessment Scale**

Table 3 presents the mean and standard deviations for RAS scores for patients who were classified as recovered compared to those who were not.

Table 3

*Descriptive Statistics for Patient Scores on the Recovery Assessment Scale who have been Classified as Recovered or Not Recovered Based on Each DASS-21 Scale Using the JT and HA Calculation Methods*

<table>
<thead>
<tr>
<th>Classifications based on JT method</th>
<th>Classifications based on HA method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>Anxiety</td>
</tr>
<tr>
<td>n</td>
<td>59</td>
</tr>
<tr>
<td>Mean</td>
<td>126.79</td>
</tr>
<tr>
<td>SD</td>
<td>14.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Not classified as recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>SD</td>
</tr>
</tbody>
</table>

Of 119 patients who completed the RAS six weeks following discharge from hospital, between 18.5% and 56.3% of patients were classified as recovered, depending on which DASS-21 scale a patient was measured on and with which clinical significance calculation method. Patients who were classified as recovered on each scale of the DASS-21 according to the JT method scored significantly higher on the RAS than those
who made no clinically significant change (Depression scale: $t(117)=4.97, p<.001, d=.92$; Anxiety scale: $t(117)=2.53, p<.05, d=.47$; Stress scale: $t(117)=4.60, p<.001, d=.85$). A similar pattern was found when patients who were classified as recovered according to the HA method were compared with those who were not (Depression scale: $t(117)=5.75, p<.001, d=1.06$; Anxiety scale: $t(117)=3.79, p<.001, d=.70$; Stress scale: $t(117)=4.27, p<.001, d=.79$). This suggests that both the JT and HA methods for evaluating clinically significant change (i.e., a classification of recovered) demonstrate construct validity, as clinically significant change on a symptom measure, the DASS-21, is related to higher scores on the RAS, representative of a more positive perception of ‘consumer-based’ recovery.

**Quality of Life**

Table 4 presents the means and standard deviations for Q-LES-Q scores for patients who were classified as recovered compared to those who were not.
Table 4

Descriptive Statistics for Patient Scores on the Quality of Life Enjoyment and Satisfaction Scale who have been Classified as “Recovered” or “Not Recovered” Based on Each DASS-21 Scale Using the JT and HA Calculation Methods

<table>
<thead>
<tr>
<th>Classified as “recovered”</th>
<th>Classifications based on JT method</th>
<th>Classifications based on HA method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Depression</td>
<td>Anxiety</td>
</tr>
<tr>
<td>n</td>
<td>413</td>
<td>264</td>
</tr>
<tr>
<td>Mean</td>
<td>58.84</td>
<td>60.79</td>
</tr>
<tr>
<td>SD</td>
<td>15.73</td>
<td>16.94</td>
</tr>
<tr>
<td>Not classified as “recovered”</td>
<td>n</td>
<td>305</td>
</tr>
<tr>
<td>Mean</td>
<td>45.03</td>
<td>48.43</td>
</tr>
<tr>
<td>SD</td>
<td>19.46</td>
<td>18.16</td>
</tr>
</tbody>
</table>

Of 718 patients who completed the Q-LES-Q at post-treatment, between 19.1% and 58.9% of patients were classified as having achieved a clinically significant change, depending on which DASS-21 scale they were assessed on, and by which clinical significance classification method. Perceived quality of life was greater for those patients classified as recovered by the JT method than those who were not (Depression scale: $t(716)=10.51, p<.001, d=.79$; Anxiety scale: $t(716)=9.01, p<.001, d=.67$; Stress scale: $t(716)=10.44, p<.001, d=.78$). Similarly, patients classified as having made a clinically significant change according to the HA method also had higher perceived quality of life (Depression scale: $t(716)=13.81, p<.001, d=1.03$; Anxiety scale: $t(716)=9.87, p<.001, d=.74$; Stress scale: $t(716)=12.37, p<.001, d=.92$). These findings provide support for the construct validity of clinically significant change as calculated
by both the JT and HA method, since a classification of clinically significant change is
associated with greater perceived quality of life.

**Readmission to Hospital within 28 Days of Discharge**

Figure 1 illustrates the percentage of patients who are readmitted to hospital within 28 days of their discharge of those who are considered recovered compared to those who are not, across the three DASS-21 scales and for each calculation method.

![Figure 1](image.png)

**Figure 1.** Proportion of patients readmitted within 28 days of discharge who are considered recovered and not recovered by the JT and HA methods used with the DASS-21 scale scores. Error bars represent 95% confidence intervals.

Figure 1 illustrates patients who are not considered to have recovered according to the calculation method and DASS-21 scale used, have a greater chance of being readmitted within 28 days of discharge. A significantly higher proportion of patients who were not considered recovered at discharge were readmitted within 28 days than those who were considered recovered by the JT method with the Depression scale ($\chi^2(1)=9.80, p=.002, \phi=.117$), the HA method with the Depression scale ($\chi^2(1)=6.93, p=.008, \phi=.098$), and the HA method with the Stress scale ($\chi^2(1)=4.259, p=.039$,
The remaining classification methods yielded no significant differences between readmission rates for recovered patients compared to non-recovered patients. Since patients who were not considered to have made a clinically significant change on the Depression scale were approximately twice as likely to be readmitted within 28 days of discharge than those patients whose change had been considered clinically significant, this provides support for the construct validity of recovery as evaluated by both calculation methods but only when classifications are based on certain DASS-21 scores.

**Comparison of Effect Sizes for JT and HA Methods**

To assess the relative size of the relationships between recovery related variables (RAS, Q-LES-Q, readmission within 28 days) and whether patients are considered to have made a clinically significant change according to the JT and HA methods, previously calculated effect sizes (Cohen’s $d$ and $\phi$) were converted to a Pearson correlation coefficient, $r$. Cohen (1988) suggests that an effect size $r$ of .10 is considered small, .30 considered medium, and .50 considered large. Table 5 presents a summary of effect sizes ($r$) and 95% confidence intervals for the relationship between each recovery variable and making a clinically significant change based on scores on each DASS-21 scale.
Table 5

Effect Sizes ($r$) and 95% Confidence Intervals (in Parentheses) for the Relationships between Making Clinically Significant Change based on DASS-21 Scale Scores and Scores on the RAS, Q-LES-Q, and Readmission within 28 Days of Discharge for Each Classification Method

<table>
<thead>
<tr>
<th></th>
<th>Depression scale</th>
<th>Anxiety scale</th>
<th>Stress scale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recovery Assessment Scale</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JT method</td>
<td>.418 (.236-.600)</td>
<td>.228 (.046-.410)</td>
<td>.391 (.209-.573)</td>
</tr>
<tr>
<td>HA method</td>
<td>.469 (.287-.651)</td>
<td>.331 (.149-.513)</td>
<td>.367 (.185-.549)</td>
</tr>
<tr>
<td><strong>Quality of Life Enjoyment and Satisfaction Questionnaire</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JT method</td>
<td>.366 (.293-.439)</td>
<td>.319 (.246-.392)</td>
<td>.363 (.290-.436)</td>
</tr>
<tr>
<td>HA method</td>
<td>.459 (.386-.532)</td>
<td>.346 (.273-.419)</td>
<td>.420 (.347-.493)</td>
</tr>
<tr>
<td><strong>Readmission within 28 days</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JT method</td>
<td>.117 (.044-.190)</td>
<td>.046 (-.027-.119)</td>
<td>.037 (-.036-.110)</td>
</tr>
<tr>
<td>HA method</td>
<td>.098 (.025-.171)</td>
<td>.052 (-.021-.125)</td>
<td>.077 (.004-.150)</td>
</tr>
</tbody>
</table>

Confidence intervals (95%) calculated around each effect size indicate that there are no significant differences between the effect sizes for each comparison made between the JT and HA methods, since they overlap in each instance. This suggests that patients who are considered recovered according to the HA method of calculation are not significantly different to those considered recovered according to the JT method, in regards to consumer-based recovery, perceived life enjoyment and satisfaction, and readmission to hospital within 28 days of discharge.

**Discussion**

It has been recommended that a measure of clinical significance be reported in all psychotherapy outcome studies (Lambert & Ogles, 2009). Although evidence exists
in the literature supporting the validity of the Jacobson-Truax method for determining clinical significance, Lambert and Ogles (2009) advised that it was necessary for an exploration of the clinical significance categories to be compared with other indices of improvement. Therefore the current study firstly aimed to explore the validity of the category recovered, which describes those patients who have made a clinically significant change during treatment. Secondly, since rates of recovery were demonstrated to be substantially higher for those patients classified using the JT method than those classified using the HA method (Ronk et al., 2012; i.e., Chapter 2), the study aimed to explore relative usefulness and meaningfulness of each calculation method’s classifications of recovery within the same sample of patients.

When patients who received a classification of recovered at post-treatment (calculated using the JT method and DASS-21 scores) were compared to those who were not considered recovered, the recovered patients had significantly higher RAS scores, indicative of a more positive consumer-based sense of recovery, and significantly higher Q-LES-Q scores, indicative of a greater perception of life enjoyment and satisfaction. The rate of hospital readmission within 28 days of discharge was significantly lower for those considered recovered according to the JT method with the Depression scale, and the HA method with the Depression and Stress scales. These findings provide further support to the growing body of literature suggesting that classifying patients as recovered according to the Jacobson-Truax method of clinical significance calculation has construct validity when used with a symptom measure.

Despite the differences in recovery rates between the more lenient, popular JT method and the more conservative, less commonly used HA method (McGlinchey et al., 2002; Ronk et al., 2012; i.e., Chapter 2), a comparison of the effect sizes yielded in each analysis did not uncover any significant differences between the two calculation methods. This suggests that there are no meaningful differences between how the two
methods capture the construct of recovery as conceptualised by the variables chosen in the current study. Therefore we echo the recommendation made by Lambert and Ogles (2009) that the Jacobson-Truax method continue to be used since it is the most commonly used and simplest to calculate. It could be argued that since the current study was correlational in nature, it was not possible to determine which method was better ‘calibrated’ towards recovery. This is true, however the issue of calibration is an arbitrary one, since the category of recovered has demonstrated meaning from the perspective of both the patient and the treatment provider. Whether the ‘true’ rate of recovered patients is indeed higher or lower than that determined by the JT method is not relevant if the arbitrary categories have meaning.

Although we can conclude here the JT and HA methods appear to have similar conceptualisations of the category of recovered, the current study does not allow for any comment to be made as to the validity of the categories of improved, unchanged, or deteriorated. Further research is required to determine the relationship between belonging in each of these categories and scores on relevant behavioural or functional indices, as well as individual patient factors. For example, it may be that patients who are unchanged during treatment have lower scores on readiness to change measures. If this is the case, then clinicians could employ specific techniques such as motivational interviewing (Rollnick, Miller & Butler, 2008) for those patients who score low on a readiness to change measure at pre-treatment, to increase their chances of making a reliable or clinically significant change during treatment.

Of particular concern to clinicians are those patients who deteriorate during treatment. As discussed by Lambert and Ogles (2009), validity studies need to focus on these patients, as they are not often included in assessments of clinical significance. One reason for their lack of inclusion in such research may be the typically low proportion of patients who receive this classification. Of course, having very few deteriorators in a
sample is desirable from a clinical perspective, but makes it more difficult to explore the correlates of deterioration, as in the current study. Since the present sample consisted of inpatients that generally score high on symptom measures, the chances of increasing symptoms enough to achieve a reliable deterioration are lower than in outpatient samples. An added complexity in regards to deteriorators is that they are not a homogenous group; the negative, reliable change required to be classified as deteriorated can occur anywhere along the range of the outcome measure. For example, a deterioration based on movement from the normal range to the mild range is qualitatively different to a deterioration based on movement from the severe range to the extremely severe range of a symptom measure. It therefore follows that the correlates of deterioration may be equally as heterogeneous. Larger samples of patients are required to meaningfully explore the correlates of this form of patient change.

Methods employed in the feedback literature (Lambert et al., 2002; Newnham & Page, 2010; Shimokawa, Lambert, & Smart, 2010) could then be used to predict which patients are “at-risk” of deteriorating, allowing clinicians to intervene mid-way through treatment.

The use of readmission to hospital within 28 days of discharge as an index of recovery has its limitations. A small proportion (approximately 5-8%) of patients who are classified as recovered are readmitted to hospital within 28 days, and not all patients who worsen (and perhaps require readmission) will be readmitted. Furthermore, patients who do require further treatment do not always require this due to the same reasons for a prior admission, nor do they always seek it from the same facility. Despite this, evaluating readmission to hospital is an objective, routinely used clinical indicator of the quality of an episode of mental health care that can provide useful information. McGlinchey et al. (2002) stated that if clinical significance classifications are valid, then they should mean something in practical terms, regarding whether an individual will
remain recovered over time. In the current sample, although the rates of readmission were lower for those patients classified as \textit{recovered} than for those who were not, being assigned this classification did not remove the possibility of readmission altogether. Future research should explore the factors associated with readmission to hospital subsequent to making a clinically significant change during the initial admission.

Since participants in the current study had diagnoses predominantly of mood and anxiety disorders, the current findings should generalise well to most psychiatric populations. However, for populations with mood and anxiety disorders, scores derived from self-report measures (e.g., Q-LES-Q) may be influenced by patients’ current mood, their level of insight, or recent life events (Atkinson, Zibin, & Chuang, 1997). This issue is present in all self-report studies in psychiatric samples, and relates also to the symptom measures on which clinically significant change is measured. Furthermore, the treatment provided to patients in the current was voluntary and provided within an inpatient setting, therefore further research may be required to explore whether the validity of clinical significance classifications is supported in those populations where treatment is involuntary, or provided in outpatient settings. Finally, the patients who responded in the current study were significantly older and had longer lengths of stay than those who did not respond; several hypotheses could explain this difference. However, since a focus of the study was upon the comparison of two methods of calculating clinical significance, the differences between respondents and non-respondents were irrelevant; the more important issue was that the same patients were included in each comparison analysis.

Classifying patient change into valid clinical significance categories following mental health treatment allows treatment providers to evaluate treatment effectiveness, provide valid feedback to patients to improve their outcomes, and allows for ongoing quality improvement. Current findings suggest that classifications of clinically
significant change made using the DASS-21 demonstrate ecological, construct validity, since classifications of *recovered* align with more positive perceptions of consumer-based recovery, greater perceived life enjoyment and satisfaction, and a lower chance of being readmitted to hospital with 28 days of discharge. These results together with validity findings in the extant literature suggest that the commonly used Jacobson-Truax method of classifying clinically significant change does exhibit validity, and therefore the recommendation that clinical significance classifications are reported in every outcome study is warranted. Additionally, there was no discernable advantage to using the HA method over the JT method, therefore the use of the simpler, JT method, is recommended. A clinical significance classification can never capture *all* that treatment providers and patients find useful, or meaningful. Nonetheless, classifications of *recovered* do translate into differences in clinically meaningful indices, therefore we can confidently recommend that clinicians strive to increase the number of patients in the category of *recovered* versus making no clinically significant change. The JT methodology provides an easy, fast, and most importantly, ecologically valid way to approximate the meaningfulness of a patients’ change.
References


Callaly, T., Trauer, T., Hyland, M., Coombs, T., & Berk, M. (2011). An examination of risk factors for readmission to acute adult mental health services within 28 days


Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing Treatment Outcome of Patients at Risk of Treatment Failure: Meta-Analytic and Mega-Analytic


Chapter Four

Chapter Four consists of a manuscript in preparation: “Which component of calculating clinically significant change more validly describes post-treatment recovery?”
Which Component of Calculating Clinically Significant Change more Validly Describes Post-Treatment Recovery?

Abstract

Support has been demonstrated for the ecological validity of clinical significance classifications however it is not yet clear which of the components of these classifications (i.e., the distribution a patient belongs to post-treatment or the magnitude of the change) is more validly related to recovery. We explored each component of clinical significance classifications as calculated using scores on the Depression Anxiety Stress Scales 21 (DASS-21) in relation to three recovery variables: Recovery Assessment Scale (Corrigan, Salzer, Ralph, Sangster, & Keck, 2004) score, Quality of Life Enjoyment and Satisfaction Questionnaire (Endicott, Nee, Harrison, & Blumenthal, 1993), and whether patients had been readmitted to hospital within 28 days of discharge. Associations between recovery variables and patient clinical significance classifications suggest that the distribution a patient belonged to post-treatment was more strongly related to recovery than the magnitude of change made on DASS-21 scales during treatment. This creates an argument for the inclusion of specifying the distribution a patient belongs to when describing them as either improved, unchanged, or deteriorated.
Clinical significance methodology provides a simple and ecologically valid way to evaluate patient outcomes following treatment as demonstrated in Chapter 3. It is calculated using a cut-off point separating the ‘dysfunctional’ population from the ‘functional’ population, and a reliable change index (RCI; see Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991). Based on these two criteria, or components, a patient change can be classified as one of four categories: recovered when the patient has made a reliable change and has moved from the dysfunctional population to the functional population; improved when the patient has made a reliable change but remains in the dysfunctional population; unchanged when the patient has not made a reliable change, and deteriorated when the patient has made a reliable change in a negative direction.

Clinical significance classifications require that these two components are calculated independently and then combined; a process different to many measures of change. Since the classifications as a whole demonstrate ecological validity, it is important to assess each component separately against relevant measures. This is especially important since, somewhat alarmingly, some studies use only the reliable change index (RCI) to evaluate patient change and consider this alone to be a meaningful measure of change (e.g., Lunnen & Ogles, 1998; Speer & Greenbaum, 1995; Wise, 2010). There are several possible reasons why the RCI alone may be preferred to evaluate patient change. Firstly, since it involves one calculation, it is simpler than using the dual criteria necessary for evaluating the clinical significance of a change. Secondly, the RCI, a continuous measure, permits examination of scores across the whole range, whereas clinical significance methodology collates scores into a set of discrete classifications. Since continuous scores typically provide richer data than categorical variables, it may be that the RCI is seen to provide a more precise assessment of change. Thirdly, the RCI when calculated using a symptom measure may
be seen to measure an aspect of treatment outcome that aligns more closely with one of clinicians’ priorities: to reduce symptom severity. Fourthly, in a similar vein to early psychological experiments in sensation in which the goal was to detect the smallest noticeable change, the RCI could be seen to provide a measure of the amount of change just noticeable to a patient or treatment provider.

While evaluating change based on RCI may align with clinicians’ treatment goals, it may not necessarily align with patients’ goals. As argued by Jacobson, Roberts, Berns, and McGlinchey (1999) most patients are not interested in whether they have made a reliable improvement (as measured by the RCI); patients undergo treatment with the aim of ending their suffering and achieving their goals. It may therefore be that patient goals align more closely with the second criterion considered when assessing clinical significance: falling into the functional range on a measure. Consistent with Jacobson’s assertion, Michalak, Kosfelder, Meier and Schulte (2003, as cited in Schulte, 2008) demonstrated that a patient’s post-treatment range, rather than the reliability of their change, better predicts the need for subsequent treatment. This finding implies that potentially rich information can be lost if clinicians are not taking into account their patients’ post-treatment range when assessing outcomes.

Although treatment providers and researchers may have preferences as to which method they use to evaluate change, a systematic exploration of whether these methods alone accurately represent patient change is required before the use of either component alone can be recommended. Therefore, the present study will explore the extent to which each component captures elements of recovery important to patients and treatment providers. The same indices used in Chapter 3 will be used in the current study to conceptualise recovery. The first of these indices is patients’ score on the Recovery Assessment Scale (RAS; Corrigan, Salzer, Ralph, Sangster, & Keck, 2004) at six weeks post-treatment. The RAS assesses “consumer-based” recovery, which
considers factors that consumers consider important in recovery such as personal confidence and hope, willingness to ask for help, goal and success orientation, reliance on others, and not being dominated by their symptoms. Secondly, quality of life is measured using the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q; Endicott et al., 1993). As argued by Gladis, Gosch, Dishuk, and Crits-Christoph (1999), mental illness affects several domains of patient’s functioning therefore it is important to assess quality of life when considering whether a change has been clinically significant. Thirdly, readmission to hospital within 28 days of discharge is used as a proxy for a poor outcome (Callaly, Trauer, Hyland, Coombs, & Berk, 2011; Hodgson, Lewis, & Boardman, 2001; Lyons et al., 1997); hence not being readmitted within 28 days is associated with a positive outcome. For a more detailed description of the three chosen indices, see Chapter 3.

The aim of the current study is to systematically examine the extent to which each component of evaluating change (RCI or post-treatment distribution) represents recovery following treatment, as operationalised by our three chosen recovery indices. It is hypothesised that since RCI is an incremental measure, it will be more strongly associated with recovery than the categorical variable of the post-treatment distribution that a patient belongs to.

**Method**

**Participants**

Participants in the current study were the same as those in Chapter 3. Participants were inpatients who attended a private psychiatric hospital in Perth, Western Australia. All patients are invited to complete pre-treatment and post-treatment measures and provide informed consent for their data to be used for research purposes. While in hospital, patients are encouraged to attend a range of group therapies including cognitive behavioural therapy, interpersonal therapy, and activity-based therapy,
conducted by clinical psychologists, occupational therapists, and nurse therapists, while receiving care from their treating nurses and psychiatrist. Patients’ mean age was 43 years ($SD=15.1$) and the mean length of stay in hospital was 17.3 days ($SD=14.7$). Patients were given primary diagnoses by their treating psychiatrists of mood disorders (56.1%), anxiety disorders (19.4%), substance abuse disorders (13.7%), psychotic disorders (5.9%), and other diagnoses (4.9%). Fifty per cent of patients were married, 33% were single, and the remainder were separated, divorced, or widowed. From a sample of 718 patients discharged from hospital between April 2011 and January 2012, 355 patients discharged in the first half of the study period were contacted six weeks following their discharge from hospital. These participants were invited to complete follow up measures, and of these, 41.1% responded, 119 with complete questionnaire data. Data from the full sample of 718 patients was available to assess readmission to hospital within 28 days of discharge, and data from the smaller sample of 119 patients was used to assess recovery indices at six weeks post-treatment.

**Measures**

*Recovery Assessment Scale.* The RAS (Corrigan, Giffort, Rashid, Leary, & Okeke, 1999; Corrigan et al., 2004) is a 41 item scale purporting to measure constructs related to recovery such as personal confidence, hope, goal orientation, symptom reduction, and reliance on others. Possible responses range from 1 (strongly disagree) to 5 (strongly agree). The scale has strong internal consistency ($\alpha=.93$) and good test-retest reliability ($\alpha=.88$; Corrigan, Giffort, Rashid, Leary, & Okeke, 1999). Good convergent validity has been demonstrated as Pearson correlations with other recovery scales range from .20 to .68 (McNaught, Caputi, Oades, & Deane, 2007). Based on a Rasch analysis by Hancock, Bundy, Honey, James, & Tamsett (2011), a 31 item version was recommended, and is used in the current study. The 31 item RAS has a score range of 31 to 155, with a higher score representing a stronger experience of recovery.
Depression Anxiety Stress Scales 21. The DASS-21 is a 21 item scale measuring depression, anxiety, and stress symptoms. Possible responses range from 0 (did not apply to me at all) to 3 (applied to me very much, or most of the time), with scores on each scale totalled and doubled so that scores range from 0 to 42 on each scale. DASS-21 scale scores demonstrate adequate internal consistency (Henry & Crawford, 2005) and the score interpretations show good convergent and discriminant validity (Page, Hooke, & Morrison, 2007).

Quality of Life Enjoyment and Satisfaction Questionnaire – Short Form (Q-LES-Q; Endicott et al., 1993). The Q-LES-Q is a 14 item self-report measure assessing quality of life, enjoyment, and satisfaction with life. Satisfaction scores range from 1 (very poor) to 5 (very good) and are added and transformed onto a scale ranging from a minimum of 0 to a maximum of 100. Higher scores indicate a greater perceived quality of life. Q-LES-Q scores demonstrate good internal consistency (>.90), good test-retest reliability (.63-.89), and good construct validity (Ritsner, Kurs, Kostizky, Ponizovsky, & Modai, 2002; Trivedi et al., 2006).

Readmission to Hospital within 28 Days of Discharge. Hospital admission records were used to determine whether patients had been readmitted to hospital within 28 days of discharge.

Procedure

Patients completed the DASS-21 and Q-LES-Q at admission and discharge as part of routine outcome monitoring, which patients give written consent to prior to admission. Patients were invited to complete the RAS and the DASS-21 via mail at six weeks post-discharge. Reliable change index calculated using Jacobson-Truax (1991) method and cut-off separating the functional from dysfunctional population calculated using cut-off C from Jacobson & Truax (1991).
Results

Descriptive Statistics

Patients’ reliable change indices were calculated based on DASS-21 scale scores from pre-treatment to post-treatment, and from pre-treatment to six weeks post-treatment using the Jacobson-Truax calculation method. Means and standard deviations for both sets of RCI scores are shown in Table 1. Change on each scale was reliable, on average, with changes on the Depression scale of the highest magnitude and changes on the Anxiety scale of the smallest magnitude. The proportions of patients who belonged to the functional population on each scale at post-treatment and six weeks post-treatment were calculated based on the cut-off values calculated in Ronk, Hooke, and Page (2012; i.e., Chapter 2) and are presented in Table 1. The majority of patients belong to the functional range at post-treatment across each scale and at both time points.
Table 1

Means and Standard Deviations (in Parentheses) for RCIs Calculated using the DASS-21, Cut-off Scores Used and Proportions of Patients in the Functional Range at Post-Treatment and Six Weeks Post-Treatment

<table>
<thead>
<tr>
<th>N</th>
<th>Depresscon</th>
<th>Anxiety</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCI pre-treatment to post-treatment</td>
<td>718</td>
<td>4.55 (3.60)</td>
<td>2.38 (2.34)</td>
</tr>
<tr>
<td>RCI pre-treatment to six weeks post-treatment</td>
<td>119</td>
<td>4.27 (4.14)</td>
<td>2.13 (2.23)</td>
</tr>
<tr>
<td>Cut-off score separating functional and dysfunctional population</td>
<td>16.88</td>
<td>10.21</td>
<td>18.82</td>
</tr>
<tr>
<td>Proportion of patients in functional distribution at post-treatment</td>
<td>718</td>
<td>71.2%</td>
<td>60.7%</td>
</tr>
<tr>
<td>Proportion of patients in functional distribution six weeks post-treatment</td>
<td>119</td>
<td>71.4%</td>
<td>67.2%</td>
</tr>
</tbody>
</table>

The mean RAS score was 119.15 (SD=18.26) with a range from 63 to 155. The mean Q-LES-Q score at post-treatment was 52.97 (SD=18.69). Across the whole sample (n=718), 64 patients were readmitted to hospital (8.9%) within 28 days of their discharge.
Recovery Assessment Scale

To explore whether the reliability of the change made on the DASS-21 during treatment is related to consumer-based recovery, correlations were performed between the RCI for each scale and RAS score at six weeks post-treatment. Consumer-based recovery as assessed by the RAS was significantly associated with scores on the Depression scale; \( r(119)=.25, p<.01 \), and Stress scale; \( r(119)=.33, p<.001 \), but no significant relationship was observed between RAS score and Anxiety scale score \( (p=.076) \). Overall these findings suggest that the magnitude of symptom change on the Depression and Stress scales align more with consumer-based recovery than Anxiety.

An independent samples \( t \)-test was conducted to assess whether the range a patient belongs to at six weeks post-treatment (i.e., functional or dysfunctional) impacts upon their own sense of recovery as measured by the RAS. Patients who fell in the dysfunctional range of the Depression scale scored significantly lower on the RAS \( (M=102.55, SD=17.61) \) than patients who fell into the functional range at six weeks post-treatment \( (M=125.79, SD=13.79) \); \( t(119)=7.65, p<.001 \). Belonging to the dysfunctional range on the Anxiety scale was also associated with a lower RAS score \( (t(119)=3.45, p<.01) \), as was belonging to the dysfunctional range on the Stress scale \( (t(119)=5.97, p<.001) \). This suggests that belonging to the functional range of a symptom scale at six weeks post-treatment is related to consumer-based recovery. Next, given that the DASS-21 has three scales, it is possible to measure the number of functional ranges of the DASS-21 that a patient belongs to. For example, if a patient’s post-treatment scores on the DASS-21 place them within the dysfunctional range on the Anxiety scale but within the functional range on the Depression and Stress scales, then the number of functional ranges this patient belongs to is two. Correlations showed that the more functional ranges a patient belongs to post-treatment, the higher their RAS score; \( r(119)=.54, p<.001 \). This indicates that the greater the number of scales on which
the patient belongs to the functional distribution on the DASS-21, the greater their consumer-based recovery score.

A multiple regression analysis was conducted to explore the relative influence of Depression scale RCI and belonging to the functional range at six weeks post-treatment on RAS scores. The model was significant; $F(2, 116)=29.12, p<.001$, and explained 33% of the variance in RAS scores. Belonging to the functional range of the DASS-21 Depression scale at six weeks post-treatment was significantly associated with RAS score ($\beta=.57, p<.001$), whereas, perhaps surprisingly, the Depression scale RCI at six weeks post-treatment was not significantly associated to RAS score ($\beta=.03, p=.73$). A similar pattern was shown for the Anxiety scale, with the association between RAS score and belonging to the functional range of the Anxiety scale ($\beta=.27, p<.01$) stronger than the association between RAS score and Anxiety scale RCI ($\beta=.19, p<.05$), however in this instance both relationships were significant. Finally, the association between RAS score and belonging to the functional range of the Stress scale ($\beta=.45, p<.001$) was stronger than the association between RAS score and Stress scale RCI ($\beta=.11, p=.22$). Therefore, belonging to the functional range on each scale of the DASS-21 is more strongly related to consumer-based recovery than the magnitude of the change made on each scale.

**Quality of Life Enjoyment and Satisfaction Questionnaire**

Correlations were performed between the RCI based on pre- to post-treatment change and Q-LES-Q scores at post-treatment. RCI was significantly associated with perceived quality of life at discharge on the Depression scale of the DASS-21; $r(718)=.28, p<.001$, the Anxiety scale; $r(718)=.19, p<.001$, and the Stress scale; $r(718)=.31, p<.001$. These findings suggest that the reliability of the change made on all scales of the DASS-21 is associated with perceived quality of life at discharge.
The impact that the range a patient belongs to at post-treatment has on their perceived quality of life was measured using an independent samples \( t \)-test. Patients belonging to the functional range of the Depression scale at post-treatment scored significantly higher on the Q-LES-Q (\( M=59.30, SD=16.02 \)) than those belonging to the dysfunctional range (\( M=37.37, SD=15.39 \); \( t(716)=16.80, p<.001 \)). Belonging to the functional range on the Anxiety scale was also associated with greater perceived quality of life at discharge (\( t(716)=13.50, p<.001 \)), as was belonging to the functional range on the Stress scale (\( t(716)=13.66, p<.001 \)).

A bivariate correlation showed that the greater the number of scales on which patients belong to the functional distribution on the DASS-21, the greater their perceived quality of life; \( r(718)=.58, p<.001 \).

Multiple regression analyses were conducted to examine the influence of Depression scale RCI and belonging to the functional range at post-treatment on perceived quality of life. The model was significant; \( F(2, 715)=143.70, p<.001 \), and explained 29% of the variance in Q-LES-Q scores. Belonging to the functional range of the DASS-21 Depression scale at post-treatment was more strongly associated with higher Q-LES-Q scores at post-treatment (\( \beta=.50, p<.001 \)) than having a higher Depression scale RCI (\( \beta=.07, p<.05 \)), however both associations were significant. Similar patterns were observed for the Anxiety and Stress scales. Belonging to the functional range on the Anxiety scale had a greater association with higher Q-LES-Q score at post-treatment (\( \beta=.41, p<.001 \)) than having a higher Anxiety scale RCI (\( \beta=.19, p<.001 \)). Similarly, belonging to the functional range on the Stress scale was more strongly associated with higher Q-LES-Q scores at post-treatment (\( \beta=.41, p<.001 \)) than having a higher Stress scale RCI (\( \beta=.15, p<.001 \)). These results suggest that where a patient lies on a scale at post-treatment is more strongly related to their perceived quality of life than the magnitude of the change they have made during treatment.
Readmission within 28 Days of Discharge

Point-biserial correlations were performed to assess the relationship between pre- to post-treatment RCI for each DASS-21 scale and whether patients were readmitted within 28 days of their discharge from hospital. No significant relationships were observed between readmission and RCI for either the Depression ($r(718)=.05, p=.175$), Anxiety ($r(718)=.007, p=.854$), or Stress scales ($r(718)=.039, p=.30$).

Belonging to the dysfunctional range of the Depression scale at post-treatment was significantly associated with readmission to hospital within 28 days ($\chi^2(1)=11.15, p<.01$), with 6.7% of patients in the functional range readmitted compared to 14.5% of patients in the dysfunctional range being readmitted. Of patients belonging to the dysfunctional range of the Stress scale, 14.6% were readmitted to hospital within 28 days compared to 6.9% of those who belonged to the functional range. This difference was significant; $\chi^2(1)=9.91, p<.01$). The same pattern existed for the Anxiety scale (11.3% readmission rate in dysfunctional range compared to 7.3% for the functional range) however the relationship was not significant ($p=.066$).

Furthermore, the greater the number of functional ranges on the scales of the DASS-21 that patients belonged to at post-treatment, the lower their chances of readmission within 28 days of discharge; $r(718)=.12, p<.01$. Of patients who belong to the dysfunctional range on all three scales of the DASS-21 at post-treatment, 14.4% are readmitted within 28 days, compared to 5.3% of patients who belong to the functional range on all three scales.

Since the relationships between RCI on each DASS-21 scale and readmission within 28 days were not significant, the relative influences of RCI and number of functional distributions a patient belonged to was not analysed.
Discussion

Results showed that the magnitude of the change made on each DASS-21 scale as measured by the RCI for all three scales was significantly related to perceived quality of life across each DASS-21 scale, and also to consumer-based recovery as measured by the RAS, but only for the Depression and Stress scales. This suggests that consumer-based recovery may not relate to physical anxiety symptoms, as captured by the Anxiety scale of the DASS-21. Although it was hypothesised that since the reliable change index was a continuous scale, it would relate more strongly to recovery than the distribution that a patient belongs to (a categorical variable), the opposite was observed. The RCI for each DASS-21 scale did not relate to readmission within 28 days of discharge. When considering a patient’s need for a readmission to hospital, it is intuitive that this would occur based on a patient’s symptoms worsening to the point where they resemble a member of the dysfunctional population, considering that the dysfunctional population represents those individuals who are seeking treatment. Therefore, it follows that the RCI should not have as much of an impact on a patient’s chances of recovery as the distribution that they belong to when they end treatment.

The distribution that a patient belonged to was significantly associated all three recovery indices investigated, with the exception of the relationship between being readmitted to hospital and belonging to the dysfunctional distribution of the Anxiety scale of the DASS-21. It is seems reasonable to expect that a patient who falls within the functional range of a particular outcome measure will report a greater sense of consumer-based recovery, a more positive perception of their quality of life, and a lower chance of being readmitted to hospital within 28 days of being discharged. When we explored the distributions further, it seemed that the more functional distributions of the three DASS-21 scales a patient belonged to, the more their responses were indicative of recovery. This suggests that the more outcome measures we can evaluate patient change.
using, the more information we receive regarding their recovery. Furthermore, the more domains of your functioning fit with what is considered ‘functional’ the less likely you are to be readmitted to hospital, the greater your sense of consumer-based recovery, and the more positive your perception of your quality of life is. It is therefore recommended that clinicians draw from as many outcome measures as possible when evaluating patient change.

The current findings have implications for clinicians in terms of treatment goals. Since belonging to the functional distribution is more representative of recovery, it should follow that the aim from a clinician’s perspective is to return patients to the functional distribution, that is, to return them to a state in which they resemble a member of the normal, non-treatment-seeking population. Making reliable change on a symptom measure during treatment is of course desirable and it does relate to recovery, but has a weaker relationship to recovery than does belonging to the functional population. Furthermore, there are many instances where clinicians are interested only in the magnitude of the symptomatic change made, as a way of measuring the effect that they have had on symptom reduction for their patient. Where this is the case, it is still legitimate to calculate the reliable change index using a symptom measure that relates to the goal of the chosen intervention. However, if a patient does not perceive that their quality of life has improved following treatment, and if other aspects of consumer-based recovery such as hope, and goals for the future are not present, should a large reduction in symptoms still be considered a meaningful change to a clinician? This is perhaps a question to consider in future research.

Given the importance of knowing which range a patient belongs to post-treatment, this presents an interesting issue to consider in terms of the categories of unchanged and deteriorated. The distinction between improved and recovered is based on which distribution a patient belongs to; the functional or the dysfunctional, and
perhaps it is for this reason that these clinical significance categories seem to
demonstrate stronger ecological validity in regards to the variables examined here.
Being classified as unchanged or deteriorated is based solely on one’s reliable change
score, and is irrespective of the distribution they belong to post-treatment. As previously
discussed, deteriorated patients are not a homogenous group. The same is true for
patients classified as unchanged; this classification can be assigned to any patient falling
anywhere along the symptom continuum. Could additional information about the range
in which an unchanged patient or a deteriorated patient falls at post-treatment strengthen
the ecological validity of these categories? Future research could explore whether
assessing patient change using both criteria simultaneously, would provide clinicians
with even richer information about patient recovery. In doing so, the original four
categories of clinical significance would become six. Patients who belong to the
functional range post-treatment would be split into those who make positive reliable
change, no reliable change, and negative reliable change, and patients who belong to the
dysfunctional range would be split into those who make positive reliable change, no
reliable change, and negative reliable change. This creates a symmetrical classification
system, which considers both criteria for all forms of change, not just for change that is
positive and reliable, as is currently the case with clinical significance classifications.
Knowing which range a patient belongs to post-treatment, combined simultaneously
with the reliability of the change made may allow for a richer classification system for
outcome researchers and treatment providers.

The current findings are limited due to the small sample of respondents utilised,
and the 41% response rate. However, this rate of response for a mail-out survey is
greater than typically expected. In addition, the current findings cannot be generalised to
other situations in which the concept of interest is not recovery. Furthermore, since only
changes observed during inpatient treatment were measured, it is not known how the current findings apply to outpatient groups.

When evaluating patient outcomes, it is vital that clinicians are mindful of the goals in doing so. This applies not only to the goals from the clinician’s perspective, but perhaps even more importantly, the goal from the patient’s perspective.
References


Zealand Journal of Psychiatry, 41(5), 450-457. doi:
10.1080/00048670701264792


10.1348/014466506X158996


10.1111/j.1468-2850.2012.01281.x

10.1080/10503300801932505


Section Three

Exploring the versatility of clinical significance methodology
Chapter Five

Chapter Five consists of a foreword followed by a published paper:

Intuitively, a patient who experiences more severe symptoms at admission (e.g., someone who is more depressed) is more likely to make a change of a greater magnitude during treatment than a patient who experiences milder symptoms at admission. This is because they have more opportunity for improvement on the scale; therefore larger, reliable changes are probable. This higher symptom improvement for patients with severe depression is consistent with the notion of regression to the mean. This phenomenon is described by Campbell and Kenny (1999) and Davis (1976) as the tendency for extreme scores on one measure to be less extreme when measured on another measure or closer to the mean for a later measurement. However, patients who have high initial symptom severity are less likely to achieve a classification of recovered, as this requires one to pass the cut-off between the functional and dysfunctional population, which severely symptomatic patients are the furthest from. A similar problem is evident for patients who experience milder symptoms at admission. In contrast to severely symptomatic patients, mildly symptomatic patients are more likely to pass the cut-off separating the functional from the dysfunctional population, as they are closer to it to begin with. However, achieving a reliable change will be more difficult, as mildly symptomatic patients have less opportunity to improve enough on the symptom scale to be considered reliable. Therefore, it is for this reason that both the reliable change index and movement into the functional distribution is required for a classification of clinically significant change.

This is depicted in Figure 1 below. In this graphical depiction of pre-treatment and post-treatment score combinations, the four possible clinical significance classifications are labelled. The parallel diagonal lines represent the limits of the reliable change index and the horizontal line represents the cut-off between the ‘functional’ and ‘dysfunctional’ distribution.
The finding that patient’s symptom severity impacts on their chances of a good outcome is not a new observation. Elkin et al. (1995) demonstrated that the initial severity of patients’ depressive symptoms predicted their response at treatment termination. Sullivan et al. (2006) showed that those who were severely depressed initially were more likely to score in the depressed range at post-treatment than those who were mildly depressed. Similarly, Blom et al. (2007) showed that the relationship between depression severity and treatment outcome was significant, and Van et al. (2008) explored this further. They demonstrated that the probability of patients remitting from depression was lower for those who had severe depression. However, those who had severe depression symptoms showed more symptom improvement.
Barnett, van der Pols, Dobson (2004) state that regression to the mean can be assessed by examining the tail ends of a distribution. If there is more change in these areas, for example, illustrated by patients with severe depression showing the largest improvements in symptoms, there is a higher chance of regression to the mean being present. This is also suggestive of a higher likelihood of measurement error, which is commonly used to explain regression to the mean (Krause, 2008).

To explore the nature of this phenomenon in relation to clinical significance in an inpatient sample, 2610 patients of varying severities of depression were classified into clinical significance categories of recovered, improved, unchanged or deteriorated based on their scores at admission and discharge on the Depression scale of the DASS-21. At each point on the Depression scale, the proportion of patients who were classified as having recovered according to the Jacobson-Truax method and the Hageman-Arrindell method was recorded and plotted below in Figure 1. A quadratic line of best fit indicates the curvilinear relationship between initial symptom severity and probability of achieving a clinically significant change.
Figure 1. Probability of achieving a classification of recovered according to the Jacobson-Truax method and Hageman-Arrindell method at each level of severity on the DASS-21 Depression scale for depressed inpatients.

As can be seen, the probability of recovery for depressed inpatients is highest for those with moderate depressive symptoms, between approximately 20 and 30 on the DASS-21 depression scale. Patients who have depressive symptoms in the severe range or the mild range have approximately a 30-40% chance of achieving a clinically significant change versus those in the moderate range who have approximately a 50-70% chance of achieving a clinically significant change. This graph illustrates a pattern that is not consistent with the notion of regression to the mean, since the highest likelihood of a classification of clinically significant change occurs in the middle of the
distribution, rather than at the tail ends. This is not surprising, since in order to make a classification of clinically significant change, one requires both a magnitude of change that is larger than that expected in the absence of a true change (RCI), and a movement from the dysfunctional to the functional distribution. However, if only the magnitude of the change is observed, that is, the RCI, then regression to the mean will be more likely to be observed.

This phenomenon of different likelihoods of recovery for patients of differing severities could potentially lead to a bias in the rates of recovery reported at different treatment sites (e.g., an inpatient clinic, where patients are generally more severely symptomatic than patients attending an outpatient clinic). This is because patients who are more unwell have a higher chance of making changes of a larger magnitude than those who are mildly unwell, but less chance of it being clinically significant. As demonstrated in the previous chapter, the range that patients belong to (dysfunctional or functional) has a larger impact on their chances of recovery than the magnitude and reliability of the change made. This provides us with a rationale to separate the dysfunctional range into two (e.g., an inpatient range and an outpatient range). This will be demonstrated in the following chapter, Chapter 5, which consists of a published paper presenting practical tools that can be used to apply clinical significance criteria to patients belonging to populations of varying severities. If clinically significant change can flex to have different meanings for patients of different levels of severity, then this will correctly imply that patients of different levels of severity have different treatment goals.

Indeed, the reliable change index also impacts on a patient’s chances of achieving a clinically significant change during treatment. Achieving reliable change is more difficult for patients with low symptom severity at commencement of treatment. However if a patient commences treatment in a range where a clinically significant change is not mathematically possible, it is assumed that a clinician will take this into account when evaluating their treatment effectiveness.
References


Krause, M. S. (2008). Regression toward the mean in effectiveness studies: Theoretically possible, not mathematically inevitable. *Quality and Quantity, 42*(6), 859-865. doi:http://dx.doi.org/10.1007/s11135-007-9137-8


Assessing Clinical Significance of Treatment Outcomes using the DASS-21

Abstract

Standard clinical significance classifications are based on movement between the ‘dysfunctional’ and ‘functional’ distributions; however this dichotomy ignores heterogeneity within the ‘dysfunctional’ population. Based on the methodology described by Tingey et al. (1996), the present study sought to present a three-distribution clinical significance model for the 21-item version of the Depression Anxiety Stress Scales (DASS-21; Lovibond & Lovibond, 1995) using data from a normative sample (n=2914), an outpatient sample (n=1000), and an inpatient sample (n=3964). DASS-21 scores were collected at pre- and post-treatment for both clinical samples and patients were classified into one of five categories based on whether they had made a reliable change and whether they had moved into a different functional range. Evidence supported the validity of the three-distribution model for the DASS-21 since inpatients who were classified as making a clinically significant change showed lower symptom severity, higher perceived quality of life and higher clinician-rated functioning than those who did not make a clinically significant change. Importantly, results suggest that the new category of recovering is an intermediate point between recovered and making no clinically significant change. Inpatients and outpatients have different treatment goals and therefore use of the concept of clinical significance needs to acknowledge differences in what constitutes a meaningful change.
Mental health practitioners ask “how successful was the treatment delivered to a particular patient?” Clinical significance calculation methods provide clinicians with a meaningful way of monitoring outcomes following treatment. They take into consideration not only whether the patient has made a change that is reliable (i.e., a change more than expected given the instrument’s measurement error), but also whether the patient has moved into the functional population. Combining these two distinct criteria permits clinicians to reflect on their patients’ outcomes in a pragmatic way, distinct from traditional outcome measurement techniques. Jacobson and Truax (1991) proposed four treatment outcome classifications based upon whether patients have met one or both criteria: recovered patients have made reliable change and moved into the functional population; improved patients have made positive reliable change but remain in the dysfunctional population; unchanged patients have not made any reliable change; and deteriorated patients have reliably worsened.

Traditionally, clinical significance categorisations have been based on “functional” and “dysfunctional” populations but this dichotomy ignores heterogeneity within the “dysfunctional” population. Two exemplar populations to which patients may belong are inpatients (i.e., those staying overnight in hospital) and outpatients (i.e., those attending appointments during the day). Due to the higher intensity and cost of care received in inpatient facilities these services tend to be reserved for people who cannot be treated in an outpatient context, often due to the severity of symptoms or the complexity of comorbid disorders. Typically, the goals for inpatients are different to those of outpatients. For inpatients, the primary goal is to become well enough to return home; even if not asymptomatic at this point. For outpatients, the primary goal may be to return to a level of functioning resembling their functioning prior to their mental health difficulties (e.g., mild to minimal symptoms). So, in calculating clinical
significance for these two distinct populations, the levels of functioning that represent ‘functional’ and ‘dysfunctional’ distributions will differ.

Awareness that different levels of patient severity need to be managed differently when calculating clinical significance is not new. Tingey, Lambert, Burlingame, and Hansen (1996) recognised that different patient groups need to be assessed against different norms. They developed a severity continuum for the General Symptom Index (GSI) of the Symptom Check-List 90-Revised (SCL-90-R; Derogatis, 1977). Patients were assessed on whether they had reliably passed one of three cut-offs, each representing the point between four adjacent and mathematically distinct distributions (asymptomatic, community, outpatient, and inpatient). The original clinical significance categories, in which two population distributions (functional and dysfunctional) are defined, suggests that patients who began treatment in the severe, dysfunctional distribution must improve to the point at which they are considered to be a part of the functional population, to be classified as recovered. As we know, this is unrealistic for many inpatients, whose primary goal is to return home, not to be asymptomatic. Tingey et al. (1996) recognised this, and considered patients who had moved reliably from one distinct population to an adjacent (and less severe) population to have made a clinically significant change. This “multiple-distribution model” allows for a potentially more realistic appraisal of patient change, given the different goals and expectations of inpatients and outpatients. Seggar, Lambert, and Hansen (2002) applied this model to the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and created three distinct groups (asymptomatic, community, and clinically symptomatic) based on collected data and multiple existing samples in the literature. However only one group consisted of clinically symptomatic patients, and therefore this model did not substantially improve upon the two-distribution method proposed by Jacobson and Truax (1991). The use of the Percentage Improvement (PI)
approach for evaluating patient change has also been put forward as a way of assessing a patient’s change against their own pre-treatment symptom severity, rather than an arbitrary amount of change (Hiller, Schindler, & Lambert, 2012). A reduction in symptom severity of more than 50% is typically regarded as a treatment response (Rush et al., 2006). However, in patients who are severely symptomatic, requiring a 50% improvement in symptoms may not recognise those who have made a clinically significant change (Rush, Thase, & Dubé, 2003).

Therefore, there is a need to develop comparable indices for other measures that include more than one clinical group. The present paper seeks to extend the approach pioneered by Tingey et al. (1996) to the 21-item Depression Anxiety Stress Scales (DASS-21; Lovibond & Lovibond, 1995) using a normative group, and two clinical groups. The DASS-21 purports to measure levels of depression, anxiety, and stress and is suitable for both clinical and non-clinical populations. This is a preliminary demonstration of this model, as we are using data from two clinical samples only; an outpatient sample from USA and an inpatient sample from Australia. There is no evidence of differences between the mean severities or the factor structure of DASS-21 scores collected in Western countries such as the United Kingdom, Canada, and Australia (see Antony, Bieling, Cox, Enns, & Swinson, 1998; Crawford, Cayley, Lovibond, Wilson, & Hartley, 2011; Henry & Crawford, 2005) therefore the contrast will be between the samples in terms of symptom severity.

The current study firstly aims to demonstrate the creation of a three-distribution model for clinical significance based on DASS-21 scores from both an inpatient and an outpatient sample. Secondly, the changes made by the current sample during treatment will be classified based on the three-distribution model. Using three-distributions will yield a new category in addition to the original four categories which we suggest be
labelled *recovered*. This category will describe patients who have made a clinically significant change (i.e., moved into an adjacent and less severe range) but have not moved into the functional range in terms of their symptoms. Third, classification rates yielded by the three-distribution model will be compared to the standard two-distribution model (Jacobson & Truax, 1991). Finally, the validity of the three-distribution model will be explored. Newnham, Harwood, and Page (2007) found that clinically significant change on the Mental Health subscales of the Medical Outcomes Short Form Questionnaire (SF-36; Brazier et al., 1992) was related to improvements in quality of life measured by the Quality of Life Enjoyment and Satisfaction Scale (Q-LES-Q; Endicott, Nee, Harrison, & Blumenthal, 1993) and symptom severity measured by the Health of the Nation Outcome Scale (HoNOS; Wing et al., 1998) in inpatients. Consequently, we expect that patients who make a clinically significant change on the DASS-21 scales (i.e., a classification of *recovered* or *recovering*) will show less dysfunction than patients who do not make a clinically significant change. Dysfunction will be measured using post-treatment scores from the mental health subscale of the SF-36, the Q-LES-Q, and the HoNOS. Furthermore, we hypothesise that the category of *recovering* will form an intermediate point separating *recovered* from those who make no clinically significant change.

**Method**

**Participants**

**Outpatient sample.** The outpatient sample consisted of 1,000 patients who attended a private outpatient behavioural and cognitive therapy clinic in New Jersey, USA, from 2008 to 2012. Data were collected throughout treatment as part of routine treatment evaluation, but for the present purposes, only pre- and post-treatment scores were used. Patients provided written consent for their data to be used for this purpose as part of their general treatment consent before their first session. Patients attending the
clinic received DSM-IV diagnoses of mood disorders (33.3%), anxiety disorders (32.9%), adjustment disorders (27.2%), and other disorders (6.6%). Patients undergo behavioural and cognitive therapies administered by psychologists, psychiatrists, clinical social workers and professional counsellors. The sample was 63.2% female, with mean age of 42.9 years (SD=16.2). The mean number of sessions attended was 8.4 (SD=7.0).

**Inpatient sample.** The inpatient sample consisted of 3,964 patients admitted to a private psychiatric facility in Western Australia from 2003 to 2008. These data were extracted from a larger database specifically for the purposes of this study. Data were collected at admission and discharge for each patient as part of an ongoing evaluation program, and written consent was obtained from each patient. Patients admitted to the clinic were diagnosed with affective disorders (66%), neurotic disorders (15.6%), substance abuse disorders (7.1%), psychotic disorders (6.1%) and other disorders including personality disorders (5.2%). While in hospital, patients were offered group therapies including interpersonal psychotherapy, cognitive behavioural therapy, and acute care programs, in addition to their psychiatric care. The sample was 73.5% female, with mean age of 41.5 years (SD=15.3). The mean length of stay was 14.9 days (SD=10.3).

**Measures**

**Depression Anxiety Stress Scales.** The DASS-21 is a 21-item self-report scale that purports to measure levels of depression, stress, and anxiety in the population. Each seven-item scale has four response options ranging from 0 (did not apply to me at all) to 3 (applied to me much, or most of the time). A maximum score of 42 (i.e., each scale is multiplied by 2 to make scores comparable to the DASS-42) on each scale indicates

---

4 Other studies that have extracted data from a similar time period include Byrne, Hooke, and Page (2010), Newnham et al. (2007), and Ronk et al. (2012; i.e., Chapter 2).
elevated depression, anxiety, or stress. The DASS-21 total scale score has excellent internal consistency (0.93; Henry & Crawford, 2005), and its score interpretations have sound construct validity (Henry & Crawford, 2005; Page, Hooke, & Morrison, 2007).

**Quality of Life Enjoyment and Satisfaction Scale.** The Q-LES-Q is a 16 item self-report scale measuring the degree of enjoyment and satisfaction experienced by respondents across various domains of daily functioning. Respondents rate their satisfaction with various aspects of their life such as “social relationships” and “mood” on a 5-point scale from 1 (very poor) to 5 (very good). A maximum score of 100 indicates high perceived quality of life. The scale scores have been demonstrated to have good test-retest reliability and high internal consistency, and its score interpretations show good construct validity (Hope, Page, & Hooke, 2009; Ritsner, Kurs, Kostizky, Ponizovsky, & Modai, 2002).

**Short Form 36 Health Survey Questionnaire (SF-36; Brazier et al., 1992).** The SF-36 is a 36-item self-report measure assessing health on eight dimensions, four of which are relevant to mental health (vitality, role limitations, social functioning, and mental health). Respondents rate items on a 6-point scale from 0 (all of the time) to 5 (none of the time). A maximum score of 100 indicates minimal impairment. Scores on the SF-36 have good internal consistency (>.85; Brazier et al., 1992), and its score interpretations show acceptable convergent and discriminant validity (Brazier et al., 1992).

**The Health of the Nation Outcome Scale** (Wing et al., 1998). The HoNOS is a 12 item clinician-rated measure assessing patient functioning across 12 domains such as “cognitive problems” and “problems with relationships”. Clinicians rate items on a 5 point scale specific to each domain. Response options range from 0 (no significant problems) to 4 (severe problems), with ratings of 2 or more indicative of a clinically
significant problem in a domain. HoNOS scores have acceptable test-retest reliability and its score interpretations have good construct validity (Wing et al., 1998).

**Clinical Significance Calculation**

Despite the existence of multiple methods of calculating clinical significance (see Lambert, Hansen, & Bauer, 2008), classifications resulting from the different methods are similar (Ronk, Hooke, & Page, 2012; Chapter 2). Given the broad similarity across methods to calculate clinically significant change, the Jacobson-Truax method (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991) was used in the present research. The present study used cut-off ‘C’ as recommended by Hsu (1996) as it uses data from the functional and dysfunctional populations. For relevant formulae see Jacobson et al. (1984) and Jacobson & Truax (1991).

**Procedure**

The DASS-21 was administered to both the inpatient and outpatient samples. Q-LES-Q, SF-36, and HoNOS data were collected for the inpatient sample only.

**Results**

**Nature of the Samples**

Information from the normative, outpatient, and inpatient samples is required to calculate cut-off points and reliable change indices for calculating clinical significance. As expected, the mean pre-treatment score for outpatients (n=1000) was less severe on each DASS-21 scale than the mean pre-treatment score for inpatients (n=3964). The distribution of pre-treatment scores for outpatients and inpatients is shown in Figure 1, clearly demonstrating the heterogeneity of the two “dysfunctional” groups. The normative sample described in the DASS manual (Lovibond & Lovibond, 1995) comprised a group of 2914 (64.2% female) university students and white and blue collar workers, ranging in age from 17 to 69 years. Participants completed the DASS as part of a general health screening program. It is important to note that this sample was deemed
representative of the population as a whole, and hence may have included some people who were seeking treatment. Descriptive statistics, cut-off scores and the differences required for reliable change for outpatients and inpatients, are presented in Table 1. These are calculated based on the standard two-distribution method for calculating clinical significance described by Jacobson & Truax (1991).
Figure 1. Distributions of pre-treatment scores for inpatients and outpatients on each scale of the 21-item Depression Anxiety Stress Scales (DASS-21; Lovibond & Lovibond, 1995).
Table 1

Information Relevant to Calculating Clinical Significance based on the Jacobson-Truax method for the DASS-21 Scales in Outpatient (n=1000) and Inpatient (n=3964) Samples

<table>
<thead>
<tr>
<th></th>
<th>Depression</th>
<th>Anxiety</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population norms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[mean (SD)]<strong>a</strong></td>
<td>6.34 (6.97)</td>
<td>4.70 (4.91)</td>
<td>10.11 (7.91)</td>
</tr>
<tr>
<td>Cronbach’s alpha<strong>b</strong></td>
<td>.96</td>
<td>.92</td>
<td>.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Outpatient sample</th>
<th>Inpatient sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-treatment score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[mean (SD)]</td>
<td>13.32 (11.10)</td>
<td>16.24 (12.79)</td>
</tr>
<tr>
<td>Post-treatment score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[mean (SD)]</td>
<td>7.35 (8.96)</td>
<td>13.23 (10.66)</td>
</tr>
<tr>
<td>Standard error of difference</td>
<td>3.16</td>
<td>3.23</td>
</tr>
<tr>
<td>Cut-off score (‘C’)</td>
<td>9.22</td>
<td>10.21</td>
</tr>
<tr>
<td>Reliable change</td>
<td>6.19</td>
<td>9.23</td>
</tr>
<tr>
<td>difference required<strong>c</strong></td>
<td>6.96</td>
<td>6.33</td>
</tr>
</tbody>
</table>

**a** Lovibond & Lovibond (1995). **b** Page et al. (2007). **c** Standard error of difference multiplied by 1.96. The difference between pre- and post-treatment score must be larger than this value in order for the change to be considered reliable.
Within each scale of the DASS-21, the normal, outpatient, and inpatient distributions were mathematically distinct. As defined in Tingey et al. (1996), samples were considered mathematically distinct if they yielded both a statistically significant \((p<.05)\) \(t\)-value (independent sample, unequal variances), and at least a moderate \((d>.5)\) effect size as calculated by Cohen’s \(d\). Distinctness scores for each scale boundary are displayed in Table 2 and demonstrate that all distributions were mathematically distinct.

**Table 2**

*Distinctness of Distributions for Each Scale of the DASS-21*

<table>
<thead>
<tr>
<th></th>
<th>Depression Scale</th>
<th>Anxiety Scale</th>
<th>Stress Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Normal vs Outpatient</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(t)(df)</td>
<td>18.66* (1279)</td>
<td>14.96* (1218)</td>
<td>14.06* (1451)</td>
</tr>
<tr>
<td>Cohen’s (d)</td>
<td>.75</td>
<td>.62</td>
<td>.54</td>
</tr>
<tr>
<td><strong>Outpatient vs Inpatient</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(t)(df)</td>
<td>47.67* (1574)</td>
<td>42.78* (1994)</td>
<td>43.10* (1593)</td>
</tr>
<tr>
<td>Cohen’s (d)</td>
<td>1.67</td>
<td>1.38</td>
<td>1.51</td>
</tr>
</tbody>
</table>

*\(p<.001\).*

**Three-Distribution Clinical Significance Model for DASS-21**

To create a three-distribution clinical significance model for the DASS-21, means, standard deviations, and reliabilities (Cronbach’s alpha) are required for each population of interest. In the current demonstration, we focus on estimates of three population distributions; a normal range (based on normative sample), an outpatient range, and an inpatient range. Cut-off points were calculated using the formula for cut-off ‘C’ (see Jacobson & Truax, 1991) and using information from the two populations.
adjacent to the cut-off, as described by Tingey et al. (1996). Reliable change indices\textsuperscript{5} were calculated for each possible movement within and between the three populations using methods described by Tingey et al. (1996). These values are presented in Table 3. Since the magnitude of change required for movement between the inpatient and normal population will always yield a statistically reliable change, these values are not presented.

\textsuperscript{5} Reliable change indices are calculated by multiplying the standard error of the difference (Sdiff) by 1.96. Sdiff = SQRT(2*((SD*SQRT(1-R))^2)), where R = reliability of scale scores. A separate Sdiff score is calculated for each possible movement within each DASS-21 scale range, using a single standard deviation score when calculating the change required for a reliable move within one range, or the mean of the two relevant standard deviations when calculating the change required for a reliable move from one range to an adjacent range.
Table 3

*Cut-off Scores and Reliable Change Indices Necessary for Classifying Patient Change Using Three-Distribution Method Based on DASS-21 Scale Scores*

<table>
<thead>
<tr>
<th>Cut-off point separating:</th>
<th>DASS-21 Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) normal range and outpatient range</td>
<td>Depression 6.27 Anxiety 6.20 Stress 6.27</td>
</tr>
<tr>
<td>(b) outpatient range and inpatient range</td>
<td>Depression 6.27 Anxiety 6.20 Stress 6.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minimum movement required for reliable change:</th>
<th>DASS-21 Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) within normal range</td>
<td>Depression 3.86 Anxiety 3.85 Stress 4.90</td>
</tr>
<tr>
<td>(b) between normal and outpatient range</td>
<td>Depression 5.01 Anxiety 5.38 Stress 5.55</td>
</tr>
<tr>
<td>(c) within outpatient range</td>
<td>Depression 6.15 Anxiety 6.92 Stress 6.20</td>
</tr>
<tr>
<td>(d) between outpatient and inpatient range</td>
<td>Depression 6.28 Anxiety 8.08 Stress 6.36</td>
</tr>
<tr>
<td>(e) within inpatient range</td>
<td>Depression 6.41 Anxiety 9.23 Stress 6.52</td>
</tr>
</tbody>
</table>

*Note. A change from the inpatient range to the normal range will always be reliable; therefore, these values are not presented here.*
In this three-distribution model of clinical significance, patients are classified into one of five categories based on their reliable change index and whether they have passed the appropriate cut-off point: (1) *recovered*, when a patient has made a positive reliable change and moved into the normal range; (2) *recovering*, when a patient has made a positive reliable change and moved into the outpatient range; (3) *improved*, when a patient has made a positive reliable change without moving into an adjacent range; (4) *unchanged*, when a patient has not made a reliable change in either direction; and (5) *deteriorated*, when a patient has made a negative reliable change. When classifying patient change, one should first determine which ranges the patient has moved within or between, and select from Table 3 the appropriate “minimum movement required for reliable change” value. If the difference between the patient’s pre-treatment score and post-treatment score exceeds this value, then the change is considered to be reliable. When a reliable change is positive, the patient will either be classified as *recovered* (if they have moved from a patient range into the normal range), *recovering* (if they have moved from the inpatient range to the outpatient range), or *improved* (if they have not passed any cut-offs).

Five patients’ pre- and post-treatment depression scale scores have been selected and plotted in Figure 2 to illustrate five different clinical significance profiles.
Patient A has remained within the inpatient range (a change from 40 to 26); however since the magnitude of the change (14 scale points) is greater than the reliable change index for movement within the inpatient range (6.41), this change is considered to be reliable. Therefore, Patient A is classified as *improved*. Patient B has moved from the inpatient range (32) to the outpatient range (14), and the magnitude of this change (18 scale points) is greater than the change required for a move from the inpatient range to the outpatient range (6.28); therefore this change is reliable. Since this patient has reliably crossed from one population into an adjacent but less severe population, this is classified as *recovering*. Patient C has reliably moved from the inpatient range (28) to the normal range (4). As this patient has made a reliable move and is now in the normal range, he or she is classified as *recovered*. Patient D has moved from the outpatient range (18) to the inpatient range (30), and the magnitude of this change exceeds that
required for a movement between the outpatient range and inpatient range (6.28). As the change is reliable, and the patient’s symptom severity has increased, the patient is classified as deteriorated. Finally, Patient E has moved from the outpatient range (12) to the normal range (8) but this change is not reliable because the magnitude of the change (4 scale points) does not exceed the reliable change index for movement between the outpatient and normal ranges (5.01). Therefore, Patient E is considered unchanged.

The present sample of inpatients and outpatients were categorised using the three-distribution clinical significance classification model. The percentages of patients classified into each category on each DASS-21 scale are shown in Table 4.

Table 4

Rates of Recovery, Recovering, Improvement, No Change, and Deterioration across DASS-21 Scales in a Sample of 4964 Patients

<table>
<thead>
<tr>
<th></th>
<th>Depression scale (%)</th>
<th>Anxiety scale (%)</th>
<th>Stress scale (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recovered</td>
<td>26.4</td>
<td>22.5</td>
<td>33.9</td>
</tr>
<tr>
<td>Recovering</td>
<td>25.2</td>
<td>14.4</td>
<td>17.2</td>
</tr>
<tr>
<td>Improved</td>
<td>17.1</td>
<td>17.2</td>
<td>11.7</td>
</tr>
<tr>
<td>Unchanged</td>
<td>26.5</td>
<td>39.6</td>
<td>31.3</td>
</tr>
<tr>
<td>Deteriorated</td>
<td>4.9</td>
<td>6.3</td>
<td>5.9</td>
</tr>
</tbody>
</table>
Classifications show that approximately half of all patients in the sample made a clinically significant change (either recovered or recoversing) during the course of their treatment. The stress scale yields the highest rate of recovery, which may be due to the less stringent cut-off separating the normal range from the outpatient range. The highest rate of no change is seen within the anxiety scale; this may be due to the larger movement required for a reliable change on this scale. To further explore the new category of recoversing, we compared the rates of classification yielded by the three-distribution method presented above, with those yielded using the original two-distribution model (Jacobson & Truax, 1991). For demonstration purposes, only the classifications made based on the Depression scale will be compared here, although the patterns yielded by the Anxiety and Stress scales are similar. Complete classification rates for all three scales are available from the corresponding author. Using the values presented in Table 1, the same sample was classified into the four original categories: recovered, improved, unchanged, and deteriorated. Patients were classified as follows: 47.0% recovered, 17.4 improved, 31.8% unchanged, and 3.8% deteriorated. Figure 3 provides a visual depiction of the difference between the two classification methods, with the new category of recoversing represented by diagonal shading.
Figure 3. Comparison of classification rates for the same group of patients (n=4964) on the Depression scale of the DASS-21 using the original two-distribution model (Jacobson & Truax, 1991) and the three-distribution model (based on Tingey et al., 1996), which includes the additional category of recovering.

The increased specificity of the three-distribution model results in patients who would have previously been classified as recovered, being classified as recovering. Rates of improvement and deterioration appear to remain stable, and fewer patients are classified as unchanged using the three-distribution model.

Validity of Three-Distribution Clinical Significance Classification Model

The creation of a three-distribution model of clinical significance allows for the emergence of a fifth category, which we have termed recovering. To test the validity of this new category, contrasts were conducted on each DASS-21 scale with the aim of
demonstrating that inpatients\textsuperscript{6} who are classified as \textit{recovering} are significantly less symptomatic than inpatients who have not made a clinically significant change, but still more symptomatic than (and contrasting with) inpatients who have been classified as \textit{recovered}. Discharge scores on the mental health scale of the SF-36, the HoNOS, and the Q-LES-Q were used to measure inpatient dysfunction. Using a Bonferroni-adjusted alpha value of 0.0018, three contrasts were conducted for each scale: (1) clinically significant change (\textit{recovered} and \textit{recovering}) versus no clinically significant change, (2) \textit{recovering} versus no clinically significant change, and (3) \textit{recovering} versus \textit{recovered}. The \(t\)-value for each contrast was significant in the expected direction with the exception of the contrast between \textit{recovered} and \textit{recovering} on the Anxiety scale for the HoNOS score; \(t(3859)=1.86, p=.063\). This was as expected given that these measures are broadly symptom-based and would therefore be expected to correlate with symptom scores on the DASS-21.

Of patients classified as \textit{recovering} on each scale, most scored in the mild and moderate ranges (Depression: 34\% mild, 56.5\% moderate, 9.5\% severe; Anxiety: 30.5\% mild, 69.5\% moderate; Stress: 29.9\% normal, 42.3\% mild, 27.8\% moderate). This is consistent with the notion that recovering patients have made a clinically significant change but have not yet returned to normal functioning. All patients classified as \textit{recovered} using the three-distribution clinical significance model fell in the normal range across all three DASS-21 scales, also consistent with the description of this category.

\textbf{Discussion}

The primary aim of the current study was to demonstrate the development of a three-distribution model for assessing the clinical significance of treatment outcomes

\textsuperscript{6} Similar validation data were not available for the outpatient sample.
using the DASS-21. Using data from samples of outpatients and inpatients, as well as a normative sample from the DASS, three distinct distributions were created and patients were classified into five categories. In addition, evidence was presented supporting the validity of the fifth category, termed *recovering*.

The comparison between the inpatients and outpatients exemplifies the differences between patient groups. It also highlights the need to monitor patient outcomes using different standards and expectations depending on which pre-treatment population they belong to. On average, outpatients in our sample were predictably less severe than inpatients when beginning treatment. The demonstration of a three-distribution model using data from normative, outpatient, and inpatient samples allowed patients to be classified into one of five categories. Clinically significant change is defined when reliable, positive movement occurs between ranges. For example, a patient who reliably moves from the outpatient to the normal range has made a clinically significant change (with a classification of *recovered*). Likewise, a patient who reliably moves from the inpatient to the outpatient range has made clinically significant change (classification of *recovering*).

It is important to note that in the current study, the labels of “inpatient range” and “outpatient range” were used to characterise levels of severity. The labels were based on the general characteristics of our inpatient and outpatient samples, but do not represent *all* members of these groups. Therefore, if an inpatient scores in the moderate range on the anxiety scale, their classification for that scale should be made based on where their score places them on the scale (in this case, in the outpatient range), despite the fact that they are attending an inpatient treatment facility. This allows for the patient’s progress to be measured against what is expected for someone of that severity.
level, rather than what is expected for someone attending a particular type of treatment facility.

There is an advantage to measuring patient progress in terms of a three-distribution model of clinical significance. When patient change is classified using the original two-distribution model, all patients, regardless of their initial symptom severity, have the same goal: to move into the “functional” range. For an inpatient experiencing extremely severe symptoms at pre-treatment, moving into the normal population and no longer requiring care at post-treatment is not always a realistic goal. Furthermore, the original two-distribution model implies that a patient cannot experience a clinically significant change while remaining in the dysfunctional population. The three-distribution model acknowledges that positive movement into an adjacent, but not necessarily normal, population is clinically significant. In so doing it provides a potentially more realistic goal for patients and treatment providers, based on their pre-treatment symptom severity. For example, it is conceivable that a classification of *recovery* is clinically meaningful from the perspective of both patients and treatment providers. For a patient admitted to hospital with severe symptoms who is classified as *recovery* at discharge, this is certainly a meaningful change, as they have improved to the point where they can be discharged from an inpatient facility. Typically such a person will require outpatient care following discharge to consolidate the gains made and continue along the path to recovery, but they can resume many of their usual activities now they have left hospital. From the perspective of a treatment provider, this movement from the inpatient range to the outpatient range is meaningful since the resources expended by this patient can now be directed to other patients in need.

In demonstrating a three-distribution model for the DASS-21, it was important that the new resulting category of *recovery* accurately described the patients that it
applied to. That is, *recovering* patients needed to represent a group of patients that were more functional than those classified as not having made a clinically significant change but were not as functional as *recovered* patients. Dysfunction was represented by scores on three measures at post-treatment (the mental health subscale of the SF-36, the Q-LES-Q, and the HoNOS). Results showed that patient scores on these measures at discharge corresponded to the intermediate category hypothesised. A comparison between classifications made based on the Depression scale of the DASS-21 showed that a substantial proportion of patients that were previously categorised as *recovered* were now classified as *recovering*. It followed that all patients classified as *recovered* using the three-distribution model fell into the normal range according to DASS Depression severity categories. As such, these patients formed a more homogenous group than those classified as *recovered* by the original two-distribution method, in which 57.6% fell into the normal range, 20.9% in the mild range, and 21.4% in the moderate range. Increased homogeneity in a category improves validity because it increases the likelihood that the goals of members of that category will align, and hence improves the accuracy with which treatment providers make decisions based on the categorisations. The increased specificity of classifications made using the three-distribution method also means that recovering patients who would have been considered recovered by the original method may have a greater chance of receiving ongoing outpatient care once they have been discharged from hospital, since a classification of *recovered* may have masked this need.

Two practical tools have been developed as a result of the current research findings to aid treatment providers to classify patient outcomes on the DASS-21 using the three-distribution model. Both tools are available online: [http://dx.doi.org/10.1037/a0033100.supp](http://dx.doi.org/10.1037/a0033100.supp) and are shown in Figure 4 and Figure 5. Figure 4 shows a printable Word file containing three blank plots representing each
DASS-21 scale. These plots allow treatment providers to manually plot patients’ pre- and post-treatment DASS-21 scores and evaluate the clinical significance of the change. If treatment providers have data available from multiple time points during treatment (e.g., pre-treatment, update, post-treatment, follow-up) this can also be plotted on each graph, presenting a fuller picture of patient progress.
Figure 4. A screenshot of the Word Document available online to manually classify patient outcomes on the DASS-21.
The second is a downloadable Excel file; an electronic alternative to classifying outcomes on paper. A screenshot of this tool is shown in Figure 5. When users enter raw pre-treatment and post-treatment DASS-21 scores into a table, the range in which the score falls on each scale is presented (irrespective of whether they have attended an outpatient or an inpatient treatment facility) along with a classification of the change into one of the five clinical significance categories. The patient change is also presented graphically.
Figure 5. A screenshot of the Excel file available online to electronically classify and plot patient outcomes on the DASS-21.
Clinical significance classifications can be used to guide treatment decisions, however it is important to emphasise that each clinical significance classification applies only to the DASS-21 scale that its calculation is based on. As Ronk and colleagues (2012; Chapter 2) have argued, categorisations of clients following treatment can differ greatly across outcome measures, and therefore clinicians need to make explicit reference to the psychological construct that a measurement instrument purports to measure. That is, if a person is recovered in terms of anxiety, this does not mean they are recovered in an absolute sense. The use of classifications produced based on the DASS-21 is advantageous since clinicians have information about where movement has occurred on not one, but three scales. Future research could explore the implications of different patterns of categorisation across these three scales.

The present study is limited in that only one sample from each population of interest was used in creating a three-distribution clinical significance model for the DASS-21. The three-sample continuum created by Seggar et al., (2002) was based on collected samples as well as data from multiple studies which improved sample generalizability. The three samples represented were asymptomatic community, community, and clinically symptomatic. Unlike the present study, they used two non-clinical samples, which allowed for an “asymptomatic” classification to be made. However the authors noted that this stringent classification was not always practical as it is not necessary for patients to be asymptomatic following treatment in order for treatment to be considered successful. Furthermore, due to a lack of distinction between the moderately and severely symptomatic patients, they were combined into one clinical sample. This means that patient change within the clinical range cannot be classified as specifically as it could be if multiple clinical samples were used. Given the evidence presented for the validity of the three-distribution model demonstrated here, future research may create such models based on other outcome measures. However it is
recommended that data be used from at least two patient samples (e.g., inpatient and outpatient) in order to increase the specificity of classifications within the clinical population.

The current study focused on addressing those patients who begin treatment in the severe, inpatient range. A limitation of this is that those patients who begin treatment in the mild to normal range remain unable to achieve *clinically significant change*. A category of change proposed by Tingey, Lambert, Burlingame, and Hansen (1996) called *reliable improvement* describes those patients who begin treatment in the functional population and have made a reliable change in a positive direction. Future research could aim to combine the proposed new categories (‘recovering’ and ‘reliably improved’) into a larger set of possible clinical significance classifications.

We have demonstrated the usefulness of a three-distribution method for classifying patient change on the DASS-21. Furthermore, the new category of *recovering* has shown evidence of validity with respect to the level of dysfunction experienced by these patients at post-treatment. Future research would do well to explore the patterns of change shown by *recovering* patients in the weeks and months following their initial treatment. Whether or not these patients move into the *recovered* range, how long it takes them to do so, and the correlates of this change, may have important implications for the way recovery is conceptualised from an outcomes monitoring perspective.
References


differentiation and measurement of depression, anxiety, and stress. *Behavior
Research and Therapy, 33*, 335–343. doi: 10.1016/0005-7967(94)00075-U

(2nd ed.).* Sydney: Psychology Foundation.

significance of responses by psychiatric inpatients to the mental health subscales
of the SF-36. *Journal of Affective Disorders, 98*(1-2), 91-97. doi:
10.1016/j.jad.2006.07.001

Depression Anxiety Stress Scales (DASS) in depressed clinical samples. *British
Journal of Clinical Psychology, 46*(3), 283-297. doi:
10.1348/014466506X158996

*Quality of Life Research, 11*(6), 553. doi: 10.1023/A:1016323009671

Significance Classifications When Calculation Methods and Outcome Measures
Differ? *Clinical Psychology: Science and Practice, 19*(2), 167-179. doi:
10.1111/j.1468-2850.2012.01281.x

to-treat depression. *Biological Psychiatry, 53*(8), 743-753.

Rush, A. J., Kraemer, H. C., Sackeim, H. A., Fava, M., Trivedi, M. H., Frank, E., ... &
remission in major depressive disorder. *Neuropsychopharmacology, 31*(9),
1841-1853.


Chapter Six consists of a manuscript in preparation: “Investigating the Stability of Clinically Significant Mental Health Outcomes”
Investigating the Stability of Clinically Significant Mental Health Outcomes

Abstract

The aim of the current study is to explore patterns of movement occurring for patients classified as *recovering* at discharge. This additional category was proposed by Ronk, Korman, Hooke, and Page (2013; i.e., Chapter 5) to describe patients who have made a meaningful change during treatment but have not yet returned to normal functioning. The current study also aims to explore general patterns of movement within and between clinical significance categories in the six weeks following treatment. One hundred and forty patients completed the DASS-21 at pre-treatment, post-treatment, and six weeks post-treatment. Treatment outcomes at post-treatment and six weeks post-treatment were classified using the clinical significance methodology described Ronk et al. (2013; i.e., Chapter 5), based on the work of Tingey, Lambert, Burlingame and Hansen (1996). The majority of patients (70.0% - 76.4%) classified as *recovering* at post-treatment either maintained this classification or moved into the *recovered* category in the subsequent six weeks. A large proportion of patients remain stable in the *unchanged* category across both time periods. Findings provide evidence for the construct validity of the newly proposed clinical significance category of *recovering* and suggest that recovery may be more usefully conceptualised as a combination of a process to be experienced and an outcome to be reached. This may allow for interventions to be put into place for patients at risk of worsening in the weeks subsequent to post-treatment.
Measuring patient change during treatment using clinical significance methodology allows treatment providers to evaluate the meaningfulness of each individual’s change, rather than the mean change of a group of patients (Jacobson & Truax, 1991). Traditionally, clinical significance classifications are made based on whether a patient has moved from the dysfunctional population to the functional population on a particular measure and whether the change during treatment is statistically reliable (Ronk, Hooke, & Page, 2012; i.e., Chapter 2). This produces four possible outcome classifications including recovered (reliable movement into the functional population), improved (reliable change but remaining in dysfunctional population), unchanged (no reliable change made), and deteriorated (reliable movement in a negative direction). When patients’ changes during treatment are evaluated using clinical significance, the proportion of patients assigned to each classification can be reported, providing a richer picture of the variation of change, than can be gleaned from mean scores alone.

Previous research suggests that changes made during therapy are generally stable at follow-up (Lambert & Bergin, 1994), yet assessing change during a follow-up period using clinical significance categories tells a more comprehensive story. Anderson and Lambert (2001) demonstrated that although mean change appeared stable during a six month follow-up period (i.e., a lack of mean differences between post-treatment and follow-up scores), when assessed using clinical significance methodology, 38 out of 51 patients (74.5%) changed classifications during the six months following treatment (e.g., making no reliable change during treatment as measured at post-treatment, but achieving reliable change in the follow-up period). Expressed differently, it is not possible to conclude on the basis of the consistency of group scores from post-treatment to follow-up that individual scores are stable from post-treatment to follow-up.
Anderson and Lambert (2001) acknowledged that achieving clinically significant change is a demanding standard, with only 38% of patients in their sample attaining this level of change. Others have demonstrated rates of clinically significant change ranging from 20% to 55%, depending on the measurement method used (Ronk, Hooke, & Page, 2012; i.e., Chapter 2). Lambert, Hansen and Bauer (2008) explained these relatively low rates of clinically significant change by suggesting that for patients who experience severe and chronic mental illness, moving to a point at which they resemble a member of the functional population is an impractical goal, yet these individuals are still capable of making meaningful change. In addition, assigning patients to either a functional or a dysfunctional range creates a large amount of heterogeneity within the dysfunctional range, especially for severely symptomatic patient groups, therefore reducing the clinical value of the dichotomy (Ronk et al., 2013; i.e., Chapter 5).

Taking into consideration these issues, Ronk, Korman, Hooke and Page (2013; i.e., Chapter 5) created a new model for assessing clinical significance on the DASS-21 based on the methodology proposed by Tingey, Lambert, Burlingame and Hansen (1996). Patients are classified based on their movement within or between three ranges: inpatient, outpatient, and normal, rather than the traditional two ranges: dysfunctional and functional. They argued that this model provided a richer evaluation of the change occurring, as it resulted in a fifth category, \textit{recovering}, to be assigned to those patients who had moved reliably from the inpatient range to the outpatient range, and considered this type of change to be clinically significant. Evidence supported the validity of this category, as patients classified as \textit{recovering} by the authors were functioning at a level resembling an intermediate range between \textit{recovered} and \textit{no clinically significant change}. Ronk et al. (2013; i.e., Chapter 5) considered the category of \textit{recovering} to be clinically meaningful, because it describes patients who may no longer require inpatient
care, but do not yet resemble a member of the non-treatment-seeking population. This provides severely symptomatic patients with an achievable goal for making clinically significant change. In addition, implicit within the term *recovering* is the notion that recovery is not an outcome to be reached, but a process; ideally continuing despite patients no longer requiring inpatient care, albeit with support from outpatient services.

Although the category of *recovering* demonstrates evidence of validity based on the level of dysfunction experienced by patients at discharge (Ronk et al., 2013; i.e., Chapter 5), the patterns of change occurring in the weeks following discharge for *recovering* patients are as yet unknown. That is, do *recovering* patients indeed move into the *recovered* category in the weeks following treatment? Therefore, the primary aim of the current study is to explore the type of movement that occurs for patients who are assigned the classification of *recovering* at discharge. In addition, the current study will provide information regarding general patterns of movement within and between clinical significance categories in the six weeks following treatment. If common patterns of change are detected, then this may provide useful information regarding the likelihood of certain outcome classifications at post-treatment. Subsequently, there is potential for appropriate measures to be put in place to prevent ‘at risk’ patients from worsening once treatment ceases, based on their post-treatment classification.

**Method**

**Participants**

Participants were 355 consecutive inpatients discharged from a private psychiatric hospital between April 2011 and August 2011. Participants completed questionnaires at admission to hospital and at discharge as part of routine quality assurance procedures. Informed consent was provided by each participant on admission to hospital, and data remains de-identified. Six weeks after discharge, participants were
invited to complete further questionnaires via mail. One hundred and forty-six participants responded (41.1%), and of these, 140 had complete questionnaire data for all three time periods (admission, discharge, six weeks post-discharge). Respondents were older ($t(353)=4.68, p<.05$) and had stayed longer in hospital ($t(353)=2.16, p<.05$) than non-respondents. The mean age of respondents was 46.9 years ($SD=15.8$ years) and 75% were female. Diagnoses included mood disorders (64.3%), anxiety disorders (15.7%), psychotic disorders (7.9%), substance abuse (7.9%) and other (4.2%). The average length of stay in hospital was 19.1 days ($SD=14.9$ days). While in hospital, patients engaged in group therapies while under the care of their treating psychiatrist.

**Measures**

**Depression Anxiety Stress Scales 21 (DASS-21; Lovibond & Lovibond, 1995).** The DASS-21 is a shortened version of the 42-item DASS and consists of three 7-item scales measuring depression, anxiety, and stress. Respondents rate each item from 0 (did not apply to me at all) to 3 (applied to me very much, or most of the time) and scores within each scale are added together and multiplied by 2 for comparability with the original 42-item scale. A maximum score of 42 on each scale indicates severe symptomatology. Scores on the DASS-21 scales have adequate reliability (Henry & Crawford, 2005) and interpretations of the scale have adequate construct validity (Henry & Crawford, 2005; Page, Hooke, & Morrison, 2007).

**Procedure**

DASS-21 scores were collected for each patient at three time periods: pre-treatment (in hospital), post-treatment (in hospital), and six weeks post-treatment (mail-out). An institutional ethical review board approved this research procedure. Change occurring between (a) pre- and post-treatment, and (b) pre-treatment and six weeks post-treatment, was classified into one of five clinical significance categories as
described by Ronk et al. (2013; i.e., Chapter 5) based on the methodology presented by Tingey et al. (1996). Paired \( t \)-tests and corresponding measure of effect size, Cohen’s \( d \), will be used to assess the significance of the pre-treatment to post-treatment change. One way ANOVA and corresponding measure of effect size, eta squared (\( \eta^2 \)) will be used to assess the relationship between DASS-21 scores at post-treatment and the stability of the outcome after six weeks.

**Results**

**Descriptives**

Descriptive statistics for scores on each DASS-21 scale at pre-treatment, post-treatment, and six weeks post-treatment are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Depression</th>
<th>Anxiety</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n )</td>
<td>( M ) (SD)</td>
<td>( M ) (SD)</td>
</tr>
<tr>
<td>Pre-treatment</td>
<td>140</td>
<td>27.66 (13.25)</td>
<td>19.00 (11.14)</td>
</tr>
<tr>
<td>Post-treatment</td>
<td>140</td>
<td>11.24 (10.11)</td>
<td>8.61 (8.55)</td>
</tr>
<tr>
<td>Six weeks post-treatment</td>
<td>140</td>
<td>12.10 (11.41)</td>
<td>9.27 (8.93)</td>
</tr>
</tbody>
</table>

Paired \( t \)-tests showed that patients significantly improved between pre-treatment and post-treatment (i.e., during their stay in hospital) on the Depression scale (\( t(139)=15.09, p<.001, d=1.28 \)), Anxiety scale (\( t(139)=12.87, p<.001, d=1.09 \)), and Stress scale (\( t(139)=14.45, p<.001, d=1.23 \)). Mean differences between post-treatment and six weeks post-treatment (i.e., during their first six weeks at home following discharge from hospital) were not significant for the Depression scale (\( t(139)=.92, p=.37 \)), and several other variables.
or Anxiety scale ($t(139) = .97, p = .33$), but patients significantly worsened based on the Stress scale ($t(139) = 2.32, p < .05, d = 0.20$). This suggests that patients’ symptoms of Depression and Anxiety generally remain stable in the weeks subsequent to their discharge, but that stress symptoms increase during this time.

To examine changes at the level of the individual, the numbers of patients classified into each clinical significance category according to the changes made from (a) pre-treatment to post-treatment, and (b) pre-treatment to six weeks post-treatment, are shown for each DASS-21 scale in Table 2.
<table>
<thead>
<tr>
<th>Classification at post-treatment</th>
<th>Classification at six weeks post-treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recovered</td>
</tr>
<tr>
<td><strong>Depression scale</strong></td>
<td>Recovered</td>
</tr>
<tr>
<td></td>
<td>Recovering</td>
</tr>
<tr>
<td></td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
</tr>
<tr>
<td></td>
<td>Deteriorated</td>
</tr>
<tr>
<td><strong>Anxiety scale</strong></td>
<td>Recovered</td>
</tr>
<tr>
<td></td>
<td>Recovering</td>
</tr>
<tr>
<td></td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
</tr>
<tr>
<td></td>
<td>Deteriorated</td>
</tr>
<tr>
<td><strong>Stress scale</strong></td>
<td>Recovered</td>
</tr>
<tr>
<td></td>
<td>Recovering</td>
</tr>
<tr>
<td></td>
<td>Improved</td>
</tr>
<tr>
<td></td>
<td>Unchanged</td>
</tr>
<tr>
<td></td>
<td>Deteriorated</td>
</tr>
</tbody>
</table>

*Note.* Shaded boxes indicate stable classifications during the six weeks post-treatment. Patients who worsened during the six weeks post-discharge fall above the diagonal shading; patients who improved their classification fall fall below the diagonal shading.
The majority of patients evaluated using each scale of the DASS-21 are considered recovered at discharge, and retain this classification six weeks post-treatment. Of patients classified as recovered at discharge, the majority remain recovered six weeks later. Similarly, of patients classified as recovering at discharge, the majority either retain this classification or improve to a classification of recovered, six weeks post-treatment. Within the Depression scale, 49.2% of patients remain stable in their classifications during the six week post-treatment period, 22.9% improve their classification during this time, and 27.9% worsen. Similar patterns are observed for the Anxiety scale (54.3% stable, 22.1% improve, 25.6% worsen) and Stress scale (48.6% stable, 21.4% improve, 30% worsen). Stability is desirable for those initially given an encouraging classification (i.e., recovered, recovering, improved) but is undesirable for those given a discouraging classification at discharge (i.e., unchanged, deteriorated). Across all three scales, patients remaining stable in a classification of unchanged between discharge and six weeks post-treatment make up a substantial proportion of patients. This is especially evident for the Anxiety scale, where 22.9% of patients remain unchanged both during their hospital stay, and during the six weeks following treatment.

Scores on each DASS-21 scale at discharge are significantly related to the stability of the outcome during the following six weeks. This pattern occurs as expected, based on the possible movement within each scale. For the Depression scale, scoring lower at discharge (reporting fewer symptoms) is associated with worsening on this scale during the subsequent six weeks \( (F(2, 137)=16.44, p<.001, \eta^2=.19) \), as there is more room to move on the scale. Similar patterns occur for Anxiety \( (F(2, 137)=11.84, p<.001, \eta^2=.15) \) and Stress \( (F(2, 137)=33.25, p<.001, \eta^2=.33) \).
**Recovering Patients**

The majority of patients who are classified as *recovering* go on to either retain this classification or move into the recovered zone. The highest rate of movement into the *recovered* zone occurs within the Stress scale, in which 43.3% of patients classified as *recovering* at post-treatment, are considered *recovered* six weeks later, followed by 42.9% for the Depression scale, and 37.2% for the Anxiety scale. Overall, between 70.0% and 76.4% of patients classified as *recovering* at discharge will either maintain this positive change or improve upon it by moving into the *recovered* range over those six weeks.

**Discussion**

The aim of the current study was to explore the movements of patients falling into the *recovering* category proposed by Ronk et al., (2013; i.e., Chapter 5) in the six weeks following discharge from an inpatient facility. In addition, we aimed to explore the stability of treatment outcomes evaluated using clinical significance methodology more generally.

As expected, while differences between mean scores for post-treatment and six weeks post-treatment were not significant for scores on the Depression and Anxiety scale (but were significant, yet small, for the Stress scale), the comparison of post-discharge classifications with six weeks post-discharge comparisons revealed a combination of stability and change. Most patients classified as *recovered* at discharge remained stable during the subsequent six weeks, suggesting that recovery achieved in hospital can be maintained in the six weeks following discharge from hospital. This encouraging finding could imply, for example, that patients were adequately equipped with symptom management skills during their hospital stay and were able to implement these successfully at home during the subsequent six weeks.
While the majority of patients who remain stable within the six weeks post-
treatment have an encouraging outcome (i.e., recovered, recovering, or improved),
many remain stable within an undesirable category (i.e., unchanged or deteriorated).
For some patients, stability within the category of unchanged may reflect treatment
resistance within a particular domain, or it could be that the patient’s symptoms were
mild at admission, and remained mild at discharge. Patients who were classified as
unchanged based on their Anxiety scale score had especially high rates of stability
within this category, with 62.7% of these patients remaining unchanged. Of these,
31.3% initially scored in the normal range for Anxiety at admission therefore they have
maintained this lack of symptoms which the desired outcome. However, 68.7% of these
patients belong to the outpatient or inpatient range at admission and remain unchanged
based on their Anxiety scale score for the duration of treatment and the subsequent six
weeks. This implies that if someone who began treatment in a dysfunctional,
symptomatic range (therefore arguably requiring treatment in this area) is unchanged at
discharge, the most likely outcome for them in the subsequent six weeks is to remain
unchanged. Practical methods to address this issue might include undertaking more
assertive discharge planning than usual. For example, these patients could be equipped
with a selection of take-home anxiety coping tools (e.g., relaxation CDs, breathing
techniques written on wallet cards), thereby making potentially useful strategies more
accessible at home.

More specifically, the nature of the category of recovering was explored. Ronk
et al. (2013; i.e., Chapter 5) proposed that this category represented patients who had
made a clinically significant change during treatment, but who did not yet resemble a
member of the normal functioning population as measured on the relevant scale. It was
proposed that recovering patients could go on to be recovered, following discharge. The
present result supports the notion that recovery is a continuing process, as 37% to 43% of recovering patients move into the recovered range in the six weeks following discharge, based on DASS-21 scale scores. A further 26% to 37% of patients maintain their recovering classification. Future research could determine the optimal time following discharge to measure change in recovering patients, and the factors that increase the likelihood of a patient moving into the recovered range during this time.

Whether or not recovery is a process or an outcome is a source of debate in the recovery literature (Davidson, Drake, Schmutte, Dinzeo, & Andres-Hyman, 2009); the current findings may lend support to both perspectives. Firstly, since patients classified as recovered are most likely to maintain this classification in the weeks following discharge, it may appear that they have reached their ‘outcome’, and maintained it thus far. Secondly, the knowledge that improvement or maintenance is the most likely outcome for patients assigned a classification of recovering at discharge, supports the perspective that recovery is an ongoing ‘process’ in which changes must be made in the home environment in order for these changes to be maintained in the long term. If we considered a combination of both perspectives in our conceptualisation of recovery, for some patients it might be that recovery is the outcome of their stay in hospital, and is maintained subsequent to this, and for others it is a process that is initiated in hospital, and then continues at home, as new skills are applied to daily life. Subsequent research could explore whether patients with particular diagnoses or certain environmental stressors experience a different process of recovery, and what patients themselves perceive to be the explanations for positive post-treatment changes.

The current study was a naturalistic exploration of clinical significance classifications and the nature of their change over time. We did not aim to explore the effectiveness of a certain treatment or the likelihood of change based on certain
diagnoses, but rather, the nature of change more generally from a clinical significance classification perspective, following inpatient treatment. The aim of the category of recovering is to recognise those patients who have made a meaningful change but do not yet resemble a member of the ‘functional’ population; rather than to be a predictor of further improvement. This study demonstrates the changeable nature of categorisations post-treatment. Future research might explore the impact of undergoing outpatient treatment on the stability of outcomes following treatment. Additionally, potential correlates of improvements and deteriorations in the weeks and months following treatment could be examined (e.g., family support, living situation, employment, willingness to undertake homework activities).

Overall, it is clear that mean scores do not capture individual changes adequately, therefore the use of clinical significance classification rates is recommended when evaluating treatment outcomes, whether at discharge or follow-up. The category of *recovering* has been demonstrated to be a dynamic one; therefore recovery should be considered a process rather than a post-treatment outcome.
References


Journal of Clinical Psychology, 46(3), 283-297. doi:
10.1348/014466506X158996

10.1111/j.1468-2850.2012.01281.x


Section Four

General Discussion
General Discussion

The increasing use of clinical significance methodology to guide clinical decisions and evaluate mental health outcomes speaks to its potential value to the field of psychotherapy research. Thus it is vital that these methodologies are able to deliver what they purport to be delivering; that is, we need to be able to trust the results they yield in order to continue recommending their use in important clinical processes.

Several implicit assumptions had been made in the literature concerning clinical significance: (a) the approach taken to assess clinical significance does not matter; (b) clinical significance classifications are valid; (c) clinical significance classifications are equally appropriate for individuals at every severity level; and (d) clinical significance methodology considers classifications as outcomes rather than processes.

The current thesis posed the question: can we trust clinical significance? In answering this question, the thesis aimed to address four important questions related to clinical significance methodology:

1. How reliable are clinical significance classifications when clinically significant change is measured differently?
2. Are classifications of recovery valid?
3. Can clinical significance classifications flex to consider different patient goals?
4. Are clinical significance categories stable following treatment?

To address these questions, a series of studies investigated the consistency and meaningfulness of clinical significance categories, before demonstrating a more individualised method of classifying clinically significant change according to patients’
symptom severity, and then assessed the stability of these categories. The specific findings from these studies and associated recommendations are outlined below.

**Findings and recommendations**

**How reliable are clinical significance classifications when clinically significant change is measured differently?**

To make a clinical significance classification, one requires firstly a method by which to calculate clinically significant change, but also an outcome measure with which to measure patient change on. Several published methods claim to capture clinically significant change quantitatively (see Bryk & Raudenbush, 1992; Hageman & Arrindell, 1999; Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991; Nunnally & Kotsch, 1983; Speer, 1992) and it appears that each method claims to be superior in some way. For example, regression to the mean describes the phenomenon where extreme scores become less extreme over time (Barnett, van der Pols, & Dobson, 2005) and some methods claim to address this issue more effectively than others. By contrast, the choice of which outcome measure to use to assess patient change is almost unlimited, and relies heavily upon what constructs clinicians would like to measure in their patients. Previous research had explored whether using different calculation methods yielded significantly different classification rates (McGlinchey, Atkins, & Jacobson, 2002; Speer & Greenbaum, 1995), and whether using different outcome measures yielded different rates (Beckstead et al., 2003). However, no research to date had explored the differences in classification rates yielded when different calculation methods and different outcome measures were used simultaneously.

Five published clinical significance calculation methods were used in Chapter 2 to evaluate the treatment outcomes of 2676 depressed adult inpatients. Each of the three outcome measures chosen was related to depressed mood in some way: the Depression
scale of the DASS-21 (measuring changes in depressive symptoms; Lovibond & Lovibond, 1995); the Mental Health subscale of the Short Form Medical Outcomes Questionnaire (measuring changes in general mental health; Brazier et al., 1992); and the Quality of Life Enjoyment and Satisfaction Questionnaire (measuring changes in quality of life; Q-LES-Q; Endicott, Nee, Harrison, & Blumenthal, 1993). Results indicated that four of the five clinical significance methods used (Jacobson-Truax (JT), Gulliksen-Lord-Novick (GLN), Nunnally-Kotsch (NK), and Edwards-Nunnally (EN) methods) yielded classification rates that were similar, but in contrast to the Hageman-Arrindell (HA) method. When recovery rates were compared between the two most different calculation methods (JT and HA methods) the rates of recovery ranged from 41.1% to 53.3%. The GLN, NK and EN methods all attempt to address problems with regression to the mean, however given that their classification rates are vastly similar to that of the JT method suggests that either (a) these methods are not successful in their attempts to address regression to the mean, or (b) that regression to the mean was not present in our sample. As stated by Krause (2008), it should not be presumed that regression to the mean is present in a data set unless there is evidence suggesting that it is present. Due to the nature of the present data, many scores are in the extreme range since patients commencing treatment generally have high symptom severity. It is unclear whether regression to the mean is a factor in this data as no control group (i.e., a patient group that did not undergo treatment) was used, since the present study was observational in nature.

The use of different outcome measures yielded substantially different rates of recovery for the same patient sample, with recovery rates at 55.1% for patients assessed using their scores on the Depression scale of the DASS-21, and only 20% for patients assessed using their scores on the Q-LES-Q. Overall, the choice of clinical significance
calculation method had far less impact on classification rates than the choice of outcome measure used.

The outcome measure used needs to depend upon what type of change is most relevant and most important for the particular patient, given their diagnosis, and treatment goals. The findings in Chapter 2 are applicable to clinical practice in that they provide clear recommendations for the use of clinical significance methodology in the future. Given the differences that can arise, and possible issues with comparisons of rates of clinically significant change across different patient groups, several variables need to be clearly reported. These include patients’ diagnoses, the outcome measure selected, the clinical significance calculation method used, and the rates of clinically significant change. These more specific recommendations need to be included as part of any recommendations included. In general, clinical significance classification rates are reliable when different calculation methods are used, except when the JT and HA methods are used. The meaningfulness of this particular difference in recovery rates, and the possible effect on validity was addressed in Chapter 3.

**Are classifications of recovery valid?**

The recommendation that clinical significance classifications are included in outcome research suggests implicitly that clinical significance classifications have validity. When this assumption is combined with particular validity findings from previous research (see Ankuta & Abeles, 1993; Lambert, Okiishi, Finch, & Johnson, 1998; Lunnen & Ogles, 1998; Ogles, Lambert, & Sawyer, 1995) it could be reasonable to assume that clinical significance categories represent what they say they represent. To address the question of whether classifications of clinically significant change specifically (i.e., classifications of recovered) were ecologically valid (i.e., representative of elements of recovery important to patients and clinicians), it was first
important to find variables that captured the concept of recovery in the most meaningful way. The foreword to Chapter 3 presented findings relating to the comparison between patients’ self-perception of their change, assessed using a single question asked only after their treatment, and evaluations of their change using pre-treatment and post-treatment self-report scores on relevant outcome measures. Since patients’ rating of their change at the end of treatment was far more positive than that of pre- and post-treatment questionnaires, it was concluded that patients’ current state influences their memory of their change. Therefore, it was vital that pre- to post-treatment measures of clinically significant change were used in assessing the validity of the classification of recovery.

In exploring whether clinical significance methodology should be trusted, a comparison needed to occur between methods that yielded different rates of recovery. Since the JT method was demonstrated to be largely similar to the GLN, EN, and NK methods in Chapter 2, only the JT method alone was compared to the substantially different HA method. Recovery rates yielded by these two methods on the DASS-21 Depression scale ranged from 41.1% (HA) to 53.3% (JT). Chapter 3 compared recovered patients according to both methods across three recovery domains: (a) their score on a valid recovery measure, the Recovery Assessment Scale (RAS; Corrigan, Salzer, Ralph, Sangster, & Keck, 2004); (b) their score on a quality of life measure (Q-LES-Q); and (c) whether patients had been readmitted to hospital within 28 days of their discharge. Results demonstrated that classifications of clinically significant change were valid, in that patients considered to be recovered had significantly higher scores on the RAS, had greater perceived quality of life, and had a lower chance of being readmitted to hospital within the 28 day period following their discharge. On the question of which method (JT or HA) produced more ecologically valid classifications
of recovery, there was no significant difference between the relationships yielded. Chapter 3 provided evidence to support the ecological validity of quantitative methods for determining clinical significance. In other words, patients classified as recovered by these methods were more likely to demonstrate meaningful differences in areas that should arguably be important to both patients and clinicians alike. Therefore, classifying patients as having made a clinically significant change has validity. In the absence of a significant difference between the two methods that yielded the most different rates of recovery in patients following treatment, it is recommended that the JT method continue to be used to evaluate clinical significance following treatment. This is due to its relative ease of use, and its popularity over the HA method means that its classifications are more comparable with previous studies’ findings. Furthermore, more journals should recommend authors report data on clinical significance, and provide clear guidelines for doing so. Even more importantly, more psychotherapy researchers and clinicians should consider the use of this methodology as an ecologically valid way to evaluate patient outcomes.

A classification of clinically significant change requires (a) a change that is reliable; and (b) a change that means a patient more closely resembles a member of the functional (not treatment-seeking) population for an outcome measure than the dysfunctional (treatment-seeking) population. Given that clinically significant change as determined by these methods showed validity, Chapter 4 aimed to explore which of these two components of specifically the JT clinical significance methodology best captured recovery. The same recovery indices were used as in Chapter 3 to capture the concept of recovery. It was assumed that the reliability of the change made, as measured by the RCI, would be more strongly related to recovery, as it is a continuous variable, as opposed to a dichotomous variable. Contrary to expectations, it appeared that the distribution that a patient belonged to on a symptom measure (DASS-21) at post-
treatment was more strongly associated with elements of recovery considered important to patients and clinicians, than the magnitude, or reliability of the change made during treatment. This aligns with assumptions made regarding patients’ goals during treatment, as stated by Jacobson, Roberts, Berns, and McGlinchey (1999). They argued that patients are more interested in ending their suffering and achieving their goals, than they are in making a statistically reliable improvement on an outcome measure. This preference aligns more closely with belonging to the functional range on a symptom measure than having a high RCI following treatment. Furthermore, the higher the number of functional ranges a patient belongs to at post-treatment, the stronger their relationship to the recovery indices. Since patients’ symptom severity at post-treatment has a stronger relationship with recovery than RCI, it is recommended that clinicians aim to shift their patients into the functional range as opposed to making the largest magnitude of change on the scale. In addition, it is recommended that clinicians draw from as many outcome measures as possible when evaluating outcomes following treatment. Of course, clinicians must remain aware that some patients will have a greater chance of reaching the functional distribution than others, based on their symptom severity when they commence treatment. This important issue was addressed in Chapter 5.

**Can clinical significance classifications flex to consider different patient goals?**

As demonstrated in the foreword to Chapter 5, depressed patients who come into treatment with a moderate level of depressive symptoms have a greater chance of achieving a clinically significant change than those patients who come into treatment with either a mild level of depression or a severe level of depression. Presumably, the same phenomenon is present in patients of any diagnosis, due to the nature of what characterises clinically significant change on a measurement instrument: a large enough
movement to be considered reliable, and a movement into the functional range. Since it is more difficult for certain patients to make a change considered clinically significant, it was deemed necessary to explore the flexibility of clinical significance classifications.

In a similar vein to the work of Tingey, Lambert, Burlingame, and Hansen (1996), who used four distributions to characterise different populations of patients who completed the SCL-90R, Chapter 5 demonstrated the use of three distributions to characterise different populations of patients on the DASS-21. Existing data from a normative sample (n=2914; Lovibond & Lovibond, 1995) was used, together with data collected from an inpatient sample (n=3964) and an outpatient sample (n=1000). Five clinical significance classifications resulted: recovered, recovering, improved, unchanged, and deteriorated. The new category recovering represented those patients who had moved reliably from one patient range to a less severe range, but did not yet resemble a member of the normal range. Therefore, patients who entered treatment in the severely symptomatic range had a greater chance of having their change considered to be clinically significant. This acknowledges that patients who enter treatment in the inpatient range and leave treatment in the outpatient range have indeed made a change that is not only clinically meaningful to them (i.e., staying overnight in hospital compared to attending weekly appointments), but they have also made a change that is in line with the goals of an inpatient treatment facility (i.e., returning patients home). Furthermore, the additional category of recovering demonstrated validity, since patients who were assigned this category scored in the intermediate range between recovered and improved on symptom measures. This is logical given the findings in Chapter 4, in which it was demonstrated that the range that a patient belongs to has a stronger relationship to the concept of recovery, than the reliability of the change made.
Given the increased richness and meaning resulting from this method of classifying patients, it was recommended that more than two ranges were used to characterise patient ranges where possible, when considering the clinical significance of patient change during treatment. This allows clinicians to stay true to that which clinically significant change should represent: the *meaningfulness* of a change. Surely, a change from one distinct patient range to another, less severe patient range must be considered meaningful.

In the original two-distribution classification method, a change of a large magnitude that occurs within the dysfunctional range is not considered clinically significant, and the patient is classified as “improved”. Conversely, a patient who changed by *the minimum magnitude to be* considered reliably changed, and passed the cut-off separating functional from dysfunctional would be classified as having made a clinically significant change. Some may argue that a patient making a smaller change than another but receiving a “superior” classification creates complexity. However, it is important to emphasise, as demonstrated in Chapter 4, that crossing the boundary between two distinct patient distributions is more meaningful than making a reliable change within the same patient distribution, even if the change made while crossing the boundary is the minimum required. The new category proposed in Chapter 5, “recovering”, does not eliminate this potential complexity, but it does reduce the possibility for a change of a large magnitude to not be considered clinically significant.

It is important to note that a patient classified as “improved” within the inpatient range has still achieved an encouraging change, but they do not yet resemble a member of the outpatient population, which has important implications for that patient’s care.

The practical tools provided in Chapter 5 are aimed at clinicians and researchers alike; they provide an easy way to plot, monitor, and evaluate patient change ideally
during treatment, but of course, at the end of treatment. Therefore, Chapter 5 provides clear evidence that clinical significance categories can, and should, be flexed where possible to capture movements between meaningful patient populations. The category name of recovering implies that this classification is a transient state, or a process. The stability of clinical significance classifications was addressed in Chapter 6.

Are clinical significance categories stable following treatment?

As discussed in both Chapter 3 and Chapter 6 of this thesis, much of the recovery literature in the past two decades has focused on debate between those who consider recovery to be a state that patients either reach or do not reach at the end of treatment (medical or outcome-based recovery; see Frank et al., 1991), and those who consider recovery to be an ongoing process or journey which continues after patients cease treatment (consumer-based recovery; see Davidson, Drake, Schmutte, Dinzeo, & Andres-Hyman, 2009; Davidson, O'Connell, Tondora, Lawless, & Evans, 2005). Generally, when recovery is considered an outcome, its measurement is based on symptom severity, and when recovery is considered a process, its measurement is based more on broader factors such as hope, having goals for the future, empowerment, opportunities for growth, and feeling in control of oneself, despite experiencing symptoms.

Given the proposed addition of a new clinical significance category, recovering, in Chapter 5, it was deemed necessary to explore the nature of this category and other clinical significance categories in the weeks following treatment, against the background of the recovery debate. Through an analysis of patients (n=140) in the six weeks following their discharge from hospital, it was found that of those patients who were classified as recovering at their time of discharge, the majority of them (70-76%) either remained in the recovering category six weeks later, or improved to move into the
recovered range. That is, patients moved from a range resembling typical outpatients, to a range resembling the non-treatment-seeking population. When clinical significance categories were examined overall, it appeared that on average, around half of all patients remained stable within the clinical significance category that they were assigned at discharge. Around 22% of patients worsened in their classifications, and approximately 28% of patients improved. These findings imply that clinical significance categories can represent treatment outcomes that remain relatively stable, or that they can represent processes. It is recommended that categories are considered part of a process, and not a fixed outcome. In doing this, clinicians may be more encouraged to put in place relapse prevention strategies, even for patients who appear recovered at post-treatment. Conversely, clinicians can remain hopeful that strategies discussed during treatment may improve the chances of recovery being achieved later, even if it is not achieved by post-treatment.

Chapter 6 also demonstrated that a large proportion of patients remain unchanged both following treatment, and six weeks following post-treatment. This is concerning considering that most patients who are unchanged at the end of treatment will remain unchanged in the subsequent six weeks. Again, this has implications related to assertive discharge planning with these patients, or the consideration of alternative treatment options.

**Limitations and future research**

The current thesis is limited in its use of the DASS-21 to evaluate patient outcomes throughout the chapters. This measure was chosen due to the relevance of the three symptom subscales of depression, anxiety, and stress, to our sample of psychiatric inpatients. For this reason, the findings relating to the DASS-21 should not be generalised to other outcome measures, particularly due to the discordance
demonstrated amongst different outcome measures. However, the general processes applied to the DASS-21 in assessing the clinical significance of patient outcomes should be applied to other outcome measures of interest in further research.

The aim of patient-focused research is to reduce the gap between research and practice by developing tools that allow research findings to be directly applied to treatment (Lutz, 2002). Our vision is that the next step of clinical significance research will facilitate therapists to do this. In addition to the suggestions for future research provided in individual chapters of this thesis, it is suggested that future research explore the concept of clinically significant patient deterioration during treatment and its application to clinical practice. The proportion of patients that deteriorate is small (e.g., 3% in Ronk et al., 2012; Chapter 2; 6.8% in Speer & Greenbaum, 1995; 8.8% in Bauer et al., 2004) in comparison to those who make a reliable change. Therefore these patients have not been a main focus of clinical significance research to date. However since these outcomes are not only undesirable for the patient and treatment provider, but also financially costly, it is important that future research explore clinically significant deterioration further.

Firstly, research should determine whether clinicians can rely on classifications of deterioration in the same way that they can rely on classifications of recovery. That is, are classifications of deterioration valid? For example, do patients who are considered deteriorated have a higher rate of readmission to hospital following their initial discharge? Do these patients remain in treatment for longer? The identification of patients who deteriorate also needs to be achieved accurately and in a timely manner. At a minimum it should occur at the termination of treatment, but ideally, before treatment ends, to allow therapists to plan treatments appropriately and intervene in real time. Researchers have begun to explore this issue (see Lambert, Whipple, Smart,
Vermeersch, Nielsen, & Hawkins, 2001; Lambert et al., 2002; Harmon, Hawkins, Lambert, Slade, & Whipple, 2005; Harmon et al., 2007) and preliminary research shows that timely identification and remediation can decrease rates of deterioration.

Given this promising finding, more research is required to determine the validity of identifying patients as being “not on track” during treatment, and their pattern of deterioration. When these individuals have been reliably identified, research should explore the best methods of intervention for those patients whose “trajectories” of progress indicate a risk of poorer outcomes. For example, research has demonstrated that providing feedback to patients who are “not on track” can improve these patients’ outcomes (Lambert, 2013; Shimokawa, Lambert, & Smart, 2010). Further investigations are warranted in relation to which methods of providing feedback to “not on track” patients and their treating clinicians are the most effective.

Finally, treatments should be “matched” to patients based on their characteristics (Cuijpers, van Straten, Bohlmeijer, Hollon, & Andersson, 2010). It is important to continue exploring the correlates of clinically significant deterioration. That is, what patient factors, therapist factors, or environmental factors are associated with patient deterioration? This could allow a patient’s deterioration “risk” to be more accurately determined at the commencement of treatment, and therefore put in place relevant interventions early, such as those described in Harmon, Hawkins, Lambert, Slade, and Whipple (2005) and Harmon et al. (2007).

**General conclusions**

To be able to determine whether a patient has made a meaningful change during therapy, or perhaps, whether the therapy has had a meaningful effect on a patient, is arguably the most important question to consider when evaluating treatment outcomes.
Clinical significance methodology provides a simple way to quantitatively classify patient outcomes into distinct categories. The current thesis explored several assumptions that had been explicitly and implicitly made regarding the use and effectiveness of clinical significance methodology with the aim of making recommendations for the continued use of clinical significance methodology if it is deemed useful, or recommendations to change in response to its limitations.

In summary, clinical significance methodology should be considered trustworthy, as long as it is used as recommended. Firstly, the classifications yielded remain generally consistent amongst calculation methods used, and for this reason, the simplest method, the Jacobson-Truax method, is recommended for use. Secondly, patients classified as recovered appear to demonstrate characteristics that are considered related to the concept of recovery. Therefore, the category of recovered is meaningful, and has ecological validity. Third, clinical significance methodology can flex in a clinically meaningful and valid way to capture changes in patients from distinct patient groups that have differing goals. Fourth, clinical significance classifications yielded are not necessarily stable over time, and therefore the recovery concept should be considered to be a process. Ultimately, clinical significance methodology provides a way for clinicians to recognise clinically meaningful change in patients, with the aim of applying this knowledge to clinical practice to improve patient outcomes.
References


of study quality and effect size. *Psychological Medicine, 40*, 211-223. doi: 10.1017/S0033291709006114


Krause, M. S. (2008). Regression toward the mean in effectiveness studies: Theoretically possible, not mathematically inevitable. *Quality and Quantity, 42*(6), 859-865. doi:http://dx.doi.org/10.1007/s11135-007-9137-8


